Carlos Ivorra Castillo

ANÁLISIS MATEMÁTICO

Si una cantidad no negativa fuera tan pequeña que resultara menor que cualquier otra dada, ciertamente no podría ser sino cero. A quienes preguntan qué es una cantidad infinitamente pequeña en matemáticas, nosotros respondemos que es, de hecho, cero. Así pues, no hay tantos misterios ocultos en este concepto como se suele creer. Esos supuestos misterios han convertido el cálculo de lo infinitamente pequeño en algo sospechoso para mucha gente. Las dudas que puedan quedar las resolveremos por completo en las páginas siguientes, donde explicaremos este cálculo.

LEONHARD EULER

Índice General

Introducción

En el siglo XVII Newton y Leibniz descubren independientemente el *análisis matemático* o *cálculo infinitesimal,* una potentísima herramienta que revolucionó el tratamiento matemático de la física y la geometría, y que más tarde impregnaría las más diversas ramas de la matemática, como la estadística o la teoría de números.

Esencialmente, el cálculo infinitesimal consistía por una parte en *analizar* o descomponer la dependencia entre varias magnitudes estudiando el comportamiento de unas al variar o *diferenciar* levemente otras (lo que constituía el *cálculo diferencial)* y por otra parte en *integrar* los resultados diferenciales para obtener de nuevo resultados globales sobre las magnitudes en consideración (el llamado *cálculo integral*).

Es difícil que un lector que no tenga ya algunas nociones de cálculo pueda entender cabalmente el párrafo anterior, pero las nuevas ideas eran aún más difíciles de entender de la pluma de sus descubridores. El primer libro de texto que se publicó con el fin de explicarlas sistemáticamente fue el "Análisis" del marqués de l'Hôpital. Veamos algunos pasajes:

> *La parte infinitamente pequeña en que una cantidad variable es aumentada o disminuida de manera continua, se llama la diferencial de esta cantidad.*

Siguiendo la notación leibniziana, L'Hôpital explica que la letra d se usa para representar uno de estos incrementos infinitamente pequeños de una magnitud, de modo que dx representa un incremento diferencial de la variable x, etc.

En ningún momento se precisa qué debemos entender por un aumento infinitamente pequeño de una cantidad, pero en compensación se presentan varias reglas para tratar con diferenciales. Por ejemplo:

> *Postúlese que dos cantidades cuya diferencia es una cantidad infinitamente pequeña pueden intercambiarse una por la otra; o bien (lo que es lo mismo) que una cantidad que está incrementada o disminuida solamente en una cantidad infinitamente menor, puede considerarse que permanece constante.*

Así, por ejemplo, si analizamos el incremento infinitesimal que experimenta un producto xy cuando incrementamos sus factores, obtenemos

$$d(xy) = (x + dx)(y + dy) - xy = x\,dy + y\,dx + dxdy = x\,dy + y\,dx,$$

donde hemos despreciado el infinitésimo doble $dxdy$ porque es infinitamente menor que los infinitésimos simples $x\,dy$ e $y\,dx$.

Es fácil imaginar que estos razonamientos infinitesimales despertaron sospechas y polémicas. Baste citar el título del panfleto que en 1734 publicó el obispo de Berkeley:

> *El analista, o discurso dirigido a un matemático infiel, donde se examina si los objetos, principios e inferencias del análisis moderno están formulados de manera más clara, o deducidos de manera más evidente, que los misterios religiosos y los asuntos de la fe.*

En esta fecha el cálculo infinitesimal tenía ya más de medio siglo de historia. La razón por la que sobrevivió inmune a estas críticas y a la vaguedad de sus fundamentos es que muchos de sus razonamientos infinitesimales terminaban en afirmaciones que no involucraban infinitésimos en absoluto, y que eran confirmados por la física y la geometría. Por ejemplo, consideremos la circunferencia formada por los puntos que satisfacen la ecuación

$$x^2 + y^2 = 25.$$

Aplicando la regla del producto que hemos "demostrado" antes al caso en que los dos factores son iguales obtenemos que $dx^2 = 2x\,dx$ e igualmente será $dy^2 = 2y\,dy$. Por otra parte, $d25 = 0$, pues al incrementar la variable x la constante 25 no se ve incrementada en absoluto. Si a esto añadimos que la diferencial de una suma es la suma de las diferenciales resulta la ecuación diferencial

$$2x\,dx + 2y\,dy = 0,$$

de donde a su vez

$$\frac{dy}{dx} = -\frac{x}{y}.$$

Esto significa que si tomamos, por ejemplo, el punto $(3,4)$ de la circunferencia e incrementamos infinitesimalmente su coordenada x, la coordenada y disminuirá en $3/4\,dx$. Notemos que esto es falso para cualquier incremento finito de la variable x, por pequeño que sea, pues si valiera para incrementos suficientemente pequeños resultaría que la circunferencia contendría un segmento de la recta

$$y - 4 = -\frac{3}{4}(x-3),$$

lo cual no es el caso. Vemos que ésta se comporta igual que la circunferencia para variaciones infinitesimales de sus variables alrededor del punto $(3,4)$, aunque difiere de ella para cualquier variación finita. La interpretación geométrica es que se trata de la recta tangente a la circunferencia por el punto $(3,4)$.

El argumento será nebuloso y discutible, pero lo aplastante del caso es que nos proporciona un método sencillo para calcular la tangente a una circunferencia por uno cualquiera de sus puntos. De hecho el método se aplica a cualquier curva que pueda expresarse mediante una fórmula algebraica razonable, lo que

supera con creces a las técnicas con las que contaba la geometría analítica antes del cálculo infinitesimal.

A lo largo del siglo XIX la matemática emprendió un proceso de fundamentación que terminó con una teoría formal donde todos los conceptos están perfectamente definidos a partir de unos conceptos básicos, los cuales a su vez están completamente gobernados por unos axiomas precisos. Las ambigüedades del cálculo infinitesimal fueron el motor principal de este proceso. En los años sesenta del siglo XX se descubrió que una delicada teoría lógica, conocida como *análisis no estándar* permite definir rigurosamente cantidades infinitesimales con las que fundamentar el cálculo a la manera de Leibniz y L'Hôpital, pero no es ése el camino habitual ni el que nosotros vamos a seguir. Lo normal es erradicar los infinitésimos de la teoría, pero no así el formalismo infinitesimal. En ocasiones los símbolos dy, dx aparecen en ciertas definiciones "en bloque", sin que se les pueda atribuir un significado independiente, como cuando se define la derivada de una función $y = y(x)$ mediante

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{y(x + \Delta x) - y(x)}{\Delta x}.$$

De este modo, el cociente de diferenciales tiene el mismo significado que para Leibniz, en el sentido de que al calcularlo obtenemos el mismo número o la misma función que él obtenía, pero con la diferencia de que ya no se trata de un cociente de diferenciales, no es un cociente de nada. La definición anterior nos permite hablar de dy/dx, pero no de dy o de dx.

No obstante se puede ir más lejos y dar una definición adecuada de dx y dy de modo que se pueda probar la equivalencia

$$\frac{dy}{dx} = f(x) \quad \Longleftrightarrow \quad dy = f(x)\, dx.$$

Es algo parecido al paso de una relación algebraica como $xy^2 = x + 4y^3$, donde x e y son, digamos, números reales indeterminados, a la misma expresión entendida como una igualdad de polinomios, donde ahora x e y son indeterminadas en un sentido matemático muy preciso. Por ejemplo, según una definición habitual del anillo de polinomios $\mathbb{R}[x, y]$, la indeterminada x es la aplicación de los pares de números naturales en $\mathbb{R}$ dada por $x(1, 0) = 1$ y $x(i, j) = 0$ para cualquier otro par, es decir, algo que en nada recuerda a "un número real indeterminado".

Al introducir las formas diferenciales muchos libros modernos insisten en recalcar que los objetos como dx son "puramente formales" —como las indeterminadas en un anillo de polinomios—, que no tienen un singificado intrínseco, sino que simplemente son objetos diseñados para que se comporten según ciertas reglas que se adaptan a las propiedades de las derivadas e integrales. Llegan incluso a perdir disculpas por lo excesivamente vacía y abstracta que resulta la teoría en torno a ellos. Explican que, pese a ello, merece la pena el esfuerzo de familiarizarse con ella porque al final se ve su gran (y sorprendente) utilidad.

En este libro insistiremos en todo momento en que las diferenciales tienen un significado intrínseco muy concreto e intuitivo, y trataremos de evidenciarlo

desde el primer momento, de modo que —sin desmerecer la profundidad de la teoría— su utilidad y buen comportamiento no resulta sorprendente en absoluto. Su interpretación no será, naturalmente la de incrementos infinitesimales, sino la de aproximaciones lineales, aceptables —al menos— en los alrededores de los puntos. Esta interpretación los mantiene en todo momento muy cerca de los hipotéticos infinitésimos en los que están inspirados.

Muchos libros de física continúan trabajando con razonamientos infinitesimales al estilo antiguo, los cuales les permiten llegar rápidamente y con fluidez a resultados importantes a cambio de sacrificar el rigor lógico. Aquí adoptaremos una posición intermedia entre los dos extremos: seremos rigurosos, pero no formalistas, daremos pruebas sin saltos lógicos, pero llegaremos a resultados enunciados de tal modo que resulten "transparentes" en la práctica, emulando así la fluidez de los razonamientos infinitesimales.

Hay un caso en que los razonamientos infinitesimales están plenamente justificados, y es cuando se trata de motivar una definición. Por ejemplo, a partir de la ley de gravitación de Newton para dos masas puntuales puede "deducirse" que el campo gravitatorio generado por una distribución continua de masa contenida en un volumen V con densidad ρ viene dado por

$$E(x) = -G \int_V \frac{\rho(y)}{\|x-y\|^3}\,(x-y)\,dy.$$

La deducción no puede considerarse una demostración matemática, pues la fórmula anterior tiene el status lógico de una definición, luego es un sinsentido tratar de demostrarla. En todo caso se podría complicar la definición sustituyéndola por otra que mostrara claramente su conexión con las masas puntuales y después probar que tal definición es equivalente a la anterior. La prueba se basaría en la posibilidad de aproximar integrales por sumas finitas y con toda seguridad sería bastante prolija. Esta opción sería absurda tanto desde el punto de vista formal (¿para qué sustituir una definición sencilla por otra complicada?) como desde el punto de vista físico (¿para qué entrar en disquisiciones ϵ–δ que acabarán donde todos sabemos que tienen que acabar?). En cambio, un argumento en términos de infinitésimos convence a cualquiera de que esta definición es justamente la que tiene que ser.[1]

Del mismo modo podemos convencernos de que el potencial gravitatorio determinado por una distribución de masa ρ debe ser

$$V(x) = -G \int_V \frac{\rho(y)}{\|x-y\|}\,dy.$$

Ahora bien, de aceptar ambos hechos tendríamos como consecuencia la relación $E = -\nabla V$, pues el potencial de un campo de fuerzas es por definición la función que cumple esto. Sin embargo esto ya no es una definición, sino una afirmación sobre dos funciones que podría ser falsa en principio y que, por consiguiente, requiere una demostración. Muchos libros de física dan por sentado

[1]A cualquiera menos a un formalista puro, quien no le encontrará sentido, pero es que, como alguien dijo, "un formalista es alguien incapaz de entender algo a menos que carezca de significado."

este hecho, incurriendo así en una laguna lógica que nosotros cubriremos. Así pues, cuando el lector encuentre en las páginas que siguen un razonamiento en términos de diferenciales deberá observar que o bien desemboca en una definición o bien está completamente avalado por teoremas previos que justifican las manipulaciones de diferenciales.

Este libro ha sido escrito siguiendo cuatro guías principales:

- Presentar los resultados más importantes del análisis matemático real. Concretamente abordamos el cálculo diferencial e integral de una y varias variables reales, las ecuaciones diferenciales ordinarias y, aunque no hay ningún capítulo dedicado específicamente a ellas, estudiamos varias ecuaciones en derivadas parciales: la ecuación de Lagrange, la de Poisson, la ecuación de ondas y las ecuaciones de Maxwell. También planteamos la ecuación del calor, si bien no entramos en su estudio. Aunque, como ya hemos dicho, nos centramos en el análisis real, estudiamos las series de potencias complejas, introduciendo en particular la exponencial y las funciones trigonométricas complejas, y a partir de la teoría de funciones harmónicas y el teorema de Stokes demostramos algunos de los resultados fundamentales sobre las funciones holomorfas (esencialmente el teorema de los residuos).

- Justificar todas las definiciones, sin caer en la falacia formalista de que la lógica nos da derecho a definir lo que queramos como queramos sin tener que dar explicaciones. Pensemos, por ejemplo, en la definición de área de una superficie. Muchos libros se limitan a definirla mediante una fórmula en términos de expresiones coordenadas, sin más justificación que la demostración de su consistencia (de que no depende del sistema de coordenadas elegido). Otros aceptan como "motivación" el teorema de cambio de variables, considerando que es natural tomar como definición de cambio de variables entre un abierto de $\mathbb{R}^n$ y un abierto en una variedad lo que entre dos abiertos de $\mathbb{R}^n$ es un teorema nada trivial. No podemos resumir nuestro enfoque en pocas líneas, pero invitamos al lector a que preste atención a la justificación de éste y muchos otros conceptos.

- Mostrar la fundamentación del cálculo infinitesimal clásico, en lugar de sustituirlo por otro cálculo moderno mucho más rígido y abstracto. Por ejemplo, a la hora de desarrollar una teoría de integración potente es imprescindible introducir la teoría de la medida abstracta y sus resultados más importantes. A ello dedicamos los capítulos VII y VIII, pero tras ello, en el capítulo siguiente, envolvemos toda esta teoría abstracta en otra mucho más elástica y natural, la teoría de formas diferenciales, que requiere a la anterior como fundamento, pero que termina por ocultarla, de modo que a partir de cierto punto es muy rara la ocasión en que se hace necesario trabajar explícitamente con las medidas y sus propiedades.

- Mostrar la aplicación y la utilidad de los resultados teóricos que presentamos. Las primeras aplicaciones tienen que ver con la geometría, pero paulatinamente van siendo desplazadas por aplicaciones a la física. En

la medida de lo posible hemos evitado presentar las aplicaciones como animales enjaulados en un zoológico, es decir, desvinculadas de sus contextos naturales, de manera que den más la impresión de anécdotas que de verdaderos éxitos del cálculo infinitesimal. En el caso de la física vamos introduciendo los conceptos fundamentales (velocidad, aceleración, fuerza, energía, etc.) según van siendo necesarios, de modo que de estas páginas podría extraerse una sucinta introducción a la física. En lo tocante a la geometría, por los motivos explicados en el segundo punto nos hemos restringido a trabajar con subvariedades de $\mathbb{R}^n$, es decir, hemos evitado la definición abstracta de variedad para tener así una interpretación natural de los espacios tangentes y su relación con la variedad. En algunos ejemplos concretos necesitamos que el lector esté familiarizado con la geometría proyectiva, la teoría de las secciones cónicas y otros puntos de la geometría pre-diferencial. Los hemos marcado con un asterisco. Ninguno de estos ejemplos es necesario para seguir el resto del libro. Uno de ellos, el del plano proyectivo, lo usamos de forma no rigurosa para ilustrar la necesidad de una definición más general de variedad, mostrando que muchos de los conceptos que definimos para una subvariedad de $\mathbb{R}^3$ son aplicables formalmente al caso del plano proyectivo, si bien la teoría de que disponemos no nos permite justificar esta aplicación.

De los puntos anteriores no debe leerse entre líneas una cierta aversión hacia el análisis abstracto. Al contrario, creemos que este libro puede ser continuado de forma natural en muchas direcciones: la teoría espectral, la teoría de distribuciones, el análisis de Fourier, el cálculo variacional, la teoría de funciones de variable compleja, la geometría diferencial y la topología general.

Por citar algunos ejemplos, nosotros probamos que el problema de Dirichlet tiene solución en una familia muy amplia de abiertos para unas condiciones de frontera dadas, pero la resolución explícita en casos concretos requiere de la transformada de Fourier, que en general se aplica a muchas otras ecuaciones en derivadas parciales. Por otra parte, la transformada de Fourier permite descomponer una onda en su espectro continuo de frecuencias. Cuando se estudia la solución de la ecuación de ondas en abiertos distintos de todo $\mathbb{R}^3$ aparecen las ondas estacionarias, que llevan al análisis espectral y, en casos particulares, a la teoría de series de Fourier o de las funciones de Bessel entre otras. Los problemas de gravitación o electromagnetismo que involucran masas y cargas puntuales o corrientes eléctricas unidimensionales pueden unificarse con los problemas que suponen distribuciones continuas de masas, cargas y corrientes a través de la teoría de distribuciones.

Tampoco nos gustaría que las comparaciones que hemos hecho con otros libros se interpreten a modo de crítica. Tan sólo queremos hacer hincapié en que nuestros objetivos son distintos a los de muchos otros libros. Somos conscientes de que nuestro propósito de justificar las definiciones más allá de una motivación más o menos dudosa nos ha llevado a seguir caminos mucho más profundos y laboriosos que los habituales, por lo que, a pesar de su carácter autocontenido en lo tocante a topología y análisis, es muy difícil que este libro sea de utilidad a un lector que no cuente ya con una cierta familiaridad con la materia. Por ello

es obvio que un libro cuya finalidad principal sea didáctica, o bien que quiera profundizar más que nosotros en física o geometría diferencial, deberá pasar por alto muchas sutilezas en las que nosotros nos hemos detenido.

Comentamos, para terminar, que al lector se le supone únicamente unos ciertos conocimientos de álgebra, especialmente de álgebra lineal, y algunas nociones elementales de geometría (salvo para los ejemplos marcados con un asterisco). Esporádicamente serán necesarios conocimientos más profundos, como para la prueba de la trascendencia de e y π, sobre todo en la de π, o al estudiar el concepto de orientación, donde para interpretar el signo del determinante de una biyección afín usamos que el grupo especial lineal de $\mathbb{R}^n$ está generado por las transvecciones. Ninguno de estos hechos se necesita después.

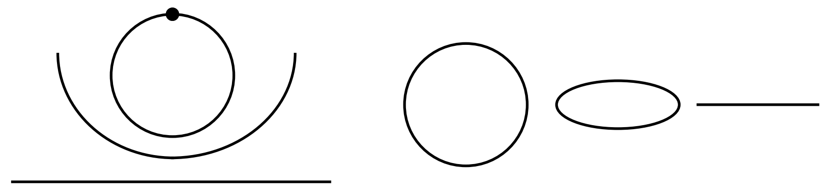
Capítulo I

Topología

La topología puede considerarse como la forma más abstracta de la geometría. El concepto principal que puede definirse a partir de la estructura topológica es el de aplicación continua, que viene a ser una transformación realizada sin cortes o saltos bruscos o, dicho de otro modo, que transforma puntos próximos en puntos próximos. Los resultados topológicos son aplicables tanto a la geometría propiamente dicha como a la descripción de otros muchos objetos más cercanos a la teoría de conjuntos general, si bien aquí nos centraremos en la vertiente geométrica. Al combinarla con el álgebra obtendremos el cálculo diferencial, que constituye la herramienta más potente para el estudio de la geometría.

1.1 Espacios topológicos

Según acabamos de comentar, una aplicación continua es una aplicación que transforma puntos próximos en puntos próximos. Nuestro objetivo ahora es definir una estructura matemática en la que esta afirmación pueda convertirse en una definición rigurosa. En primer lugar conviene reformularla así: una aplicación continua es una aplicación que transforma los puntos de alrededor de un punto dado en puntos de alrededor de su imagen. En efecto, si cortamos una circunferencia por un punto P para convertirla en un segmento, la transformación no es continua, pues los puntos de alrededor de P se transforman unos en los puntos de un extremo del segmento y otros en los puntos del otro extremo, luego no quedan todos alrededor del mismo punto. En cambio, podemos transformar continuamente (aunque no biyectivamente) una circunferencia en un segmento sin más que aplastarla.



Una forma de dar rigor al concepto de "puntos de alrededor" de un punto dado es a través de una distancia. Veremos que no es lo suficientemente general, pero sí muy representativa. La formalización algebraica de la geometría euclídea se lleva a cabo a través de $\mathbb{R}^n$. Su estructura vectorial permite definir los puntos, rectas, planos, etc. y a ésta hay que añadirle la estructura métrica derivada del producto escalar:

$$xy = \sum_{i=1}^{n} x_i\, y_i.$$

A partir de él se definen los dos conceptos fundamentales de la geometría métrica: la longitud de un vector y el ángulo entre dos vectores. En efecto, la longitud de un vector es la norma

$$\|x\| = \sqrt{xx} = \sqrt{\sum_{i=1}^{n} x_i^2},$$

y el ángulo α que forman dos vectores no nulos x, y viene dado por

$$\cos\alpha = \frac{xy}{\|x\|\,\|y\|}.$$

Estas estructuras son demasiado particulares y restrictivas desde el punto de vista topológico. La medida de ángulos es un sinsentido en topología, y la de longitudes tiene un interés secundario, pues no importan las medidas concretas sino tan sólo la noción de proximidad. En primer lugar generalizaremos el concepto de producto escalar para admitir como tal a cualquier aplicación que cumpla unas mínimas propiedades:

Definición 1.1 Usaremos la letra $\mathbb{K}$ para referirnos indistintamente al cuerpo $\mathbb{R}$ de los números reales o al cuerpo $\mathbb{C}$ de los números complejos. Si $\alpha \in \mathbb{K}$, la notación $\bar{\alpha}$ representará al conjugado de α si $\mathbb{K} = \mathbb{C}$ o simplemente $\bar{\alpha} = \alpha$ si $\mathbb{K} = \mathbb{R}$. Si H es un $\mathbb{K}$-espacio vectorial, un *producto escalar* en H es una aplicación $\cdot : H \times H \longrightarrow \mathbb{K}$ que cumple las propiedades siguientes:

a) $x \cdot y = \overline{y \cdot x}$,

b) $(x + y) \cdot z = x \cdot z + y \cdot z$,

c) $(\alpha x) \cdot y = \alpha(x \cdot y)$,

d) $x \cdot x \geq 0$ y $x \cdot x = 0$ si y sólo si $x = 0$,

para todo x, y, $z \in H$ y todo $\alpha \in \mathbb{K}$.

Notar que a) y b) implican también la propiedad distributiva por la derecha: $x \cdot (y + z) = x \cdot y + x \cdot z$.

Un *espacio prehilbertiano* es un par $(H, \cdot)$, donde H es un $\mathbb{K}$-espacio vectorial y $\cdot$ es un producto escalar en H. En la práctica escribiremos simplemente H en lugar de $(H, \cdot)$.

Si H es un espacio prehilbertiano, definimos su *norma* asociada como la aplicación $\|\ \ \| : H \longrightarrow \mathbb{R}$ dada por $\|x\| = \sqrt{x \cdot x}$.

Ejemplo Un producto escalar en el espacio $\mathbb{K}^n$ viene dado por

$$x \cdot y = x_1 \bar{y}_1 + \cdots + x_n \bar{y}_n.$$

De este modo, $\|x\| = \sqrt{|x_1|^2 + \cdots + |x_n|^2}$. ■

Teorema 1.2 *Sea H un espacio prehilbertiano y sean $x,\ y \in H$. Entonces*

a) (desigualdad de Schwarz) $|x \cdot y| \leq \|x\|\,\|y\|$.

b) (desigualdad triangular) $\|x + y\| \leq \|x\| + \|y\|$.

DEMOSTRACIÓN: a) Sean $A = \|x\|^2$, $B = |x \cdot y|$ y $C = \|y\|^2$. Existe un número complejo α tal que $|\alpha| = 1$ y $\alpha(y \cdot x) = B$. Para todo número real r se cumple

$$0 \leq (x - r\alpha y) \cdot (x - r\alpha y) = x \cdot x - r\alpha(y \cdot x) - r\bar{\alpha}(x \cdot y) + r^2\, y \cdot y.$$

Notar que $\bar{\alpha}(x \cdot y) = \bar{B} = B$, luego $A - 2Br + Cr^2 \geq 0$. Si $C = 0$ ha de ser $B = 0$, o de lo contrario la desigualdad sería falsa para r grande. Si $C > 0$ tomamos $r = B/C$ y obtenemos $B^2 \leq AC$.

b) Por el apartado anterior:

$$\begin{aligned} \|x + y\|^2 &= (x + y) \cdot (x + y) = x \cdot x + x \cdot y + y \cdot x + y \cdot y \\ &\leq \|x\|^2 + 2\|x\|\,\|y\| + \|y\|^2 = (\|x\| + \|y\|)^2. \end{aligned}$$

Notar que $x \cdot y + y \cdot x$ es un número real, luego

$$x \cdot y + y \cdot x \leq |x \cdot y + y \cdot x| \leq |x \cdot y| + |y \cdot x|.$$

■

La norma permite definir una distancia entre puntos con la que formalizar el concepto de proximidad que nos interesa, pero para ello no es necesario que la norma provenga de un producto escalar. Conviene aislar las propiedades de la norma que realmente nos hacen falta para admitir como tales a otras muchas aplicaciones:

Definición 1.3 Si E es un espacio vectorial sobre $\mathbb{K}$, una *norma* en E es una aplicación $\|\ \| : E \longrightarrow [0, +\infty[$ que cumpla las propiedades siguientes:

a) $\|v\| = 0$ si y sólo si $v = 0$.

b) $\|v + w\| \leq \|v\| + \|w\|$.

c) $\|\alpha\, v\| = |\alpha|\,\|v\|$,

para $v,\ w \in E$ y todo $\alpha \in \mathbb{K}$.

Un *espacio normado* es un par $(E, \| \ \|)$ en estas condiciones. En la práctica escribiremos E, sin indicar explícitamente la norma.

Es inmediato comprobar que la norma de un espacio prehilbertiano es una norma en el sentido general de la definición anterior. En particular $\mathbb{K}^n$ es un espacio normado con la norma del ejemplo anterior, que recibe el nombre de *norma euclídea.* El teorema siguiente nos da otras dos normas alternativas. La prueba es elemental.

Teorema 1.4 $\mathbb{K}^n$ *es un espacio normado con cualquiera de estas normas:*

$$\|x\|_1 = \sum_{i=1}^n |x_i|, \quad \|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}, \quad \|x\|_\infty = \text{máx}\big\{|x_i| \bigm| i = 1, \ldots, n\big\}.$$

Notar que para $n = 1$ las tres normas coinciden con el valor absoluto. El hecho de que estas tres aplicaciones sean normas permite obtener un resultado más general:

Teorema 1.5 *Sean $E_1, \ldots, E_n$ espacios normados. Entonces las aplicaciones siguientes son normas en $E = E_1 \times \cdots \times E_n$.*

$$\|x\|_1 = \sum_{i=1}^n \|x_i\|, \quad \|x\|_2 = \sqrt{\sum_{i=1}^n \|x_i\|^2}, \quad \|x\|_\infty = \text{máx}\big\{\|x_i\| \bigm| i = 1, \ldots, n\big\}.$$

Además se cumplen las relaciones: $\|x\|_\infty \leq \|x\|_2 \leq \|x\|_1 \leq n\|x\|_\infty$.

DEMOSTRACIÓN: Tenemos $\|x\|_i = \|(\|x_1\|, \ldots, \|x_n\|)\|_i$ para $i = 1, 2, \infty$. Usando el teorema anterior se ve inmediatamente que son normas.

$$\|x\|_\infty = \sqrt{\|x\|_\infty^2} \leq \sqrt{\sum_{i=1}^n \|x_i\|^2} = \|x\|_2 \leq \sqrt{\sum_{i=1}^n \|x_i\|^2 + 2\sum_{i<j} \|x_i\| \, \|x_j\|}$$

$$= \sqrt{\left(\sum_{i=1}^n \|x_i\|\right)^2} = \|x\|_1 \leq \sum_{i=1}^n \|x\|_\infty = n\|x\|_\infty.$$

■

Notemos también que las normas del teorema 1.4 coinciden con las construidas mediante este último teorema a partir del valor absoluto en $\mathbb{K}$.

Ejercicio: Probar que en un espacio normado se cumple $\big|\|x\| - \|y\|\big| \leq \|x - y\|$.

Como ya hemos comentado, desde un punto de vista topológico el único interés de las normas es que permiten definir la distancia entre dos puntos como $d(x, y) = \|x - y\|$. Sin embargo, a efectos topológicos no es necesario que una distancia esté definida de este modo.

Definición 1.6 Una *distancia* o *métrica* en un conjunto M es una aplicación $d : M \times M \longrightarrow [0, +\infty[$ que cumpla las propiedades siguientes

a) $d(x, y) = 0$ si y sólo si $x = y$,

b) $d(x, y) = d(y, x)$,

c) $d(x, z) \leq d(x, y) + d(y, z)$,

para todos los x, y, $z \in M$.

Un *espacio métrico* es un par (M, d) donde M es un conjunto y d una distancia en M. Como en el caso de espacios normados escribiremos M en lugar de (M, d).

Todo espacio normado E es un espacio métrico con la distancia definida por $d(x, y) = \|x - y\|$. Las propiedades de la definición de norma implican inmediatamente las de la definición de distancia. En particular en $\mathbb{K}^n$ tenemos definidas tres distancias:

$$d_1(x, y) = \sum_{i=1}^{n} |x_i - y_i|, \quad d_2(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

$$d_\infty(x, y) = \text{máx}\big\{|x_i - y_i| \bigm| 1 \leq i \leq n\big\}.$$

Más en general, estas fórmulas permiten definir distancias en cualquier producto finito de espacios métricos. La prueba del teorema siguiente es muy sencilla a partir de los teoremas 1.4 y 1.5.

Teorema 1.7 *Sean $M_1, \ldots, M_n$ espacios métricos. Sea $M = M_1 \times \cdots \times M_n$. Entonces las aplicaciones d_1, d_2, $d_\infty : M \times M \longrightarrow [0, +\infty[$ definidas como sigue son distancias en M:*

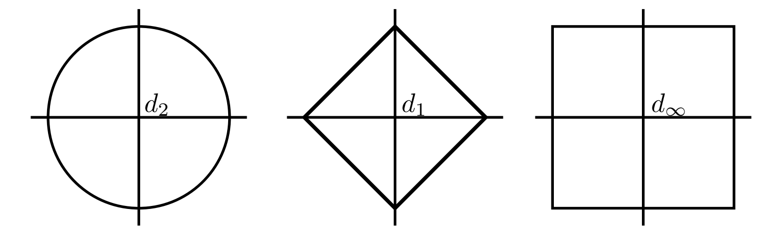
$$\begin{aligned} d_1(x, y) &= \sum_{i=1}^{n} d(x_i, y_i), \\ d_2(x, y) &= \sqrt{\sum_{i=1}^{n} d(x_i, y_i)^2}, \\ d_\infty(x, y) &= \text{máx}\{d(x_i, y_i) \mid 1 \leq i \leq n\}. \end{aligned}$$

Además se cumplen las relaciones $d_\infty(x, y) \leq d_2(x, y) \leq d_1(x, y) \leq n\, d_\infty(x, y)$.

Definición 1.8 Sea M un espacio métrico, $x \in M$ y $\epsilon > 0$ (en estos casos sobreentenderemos $\epsilon \in \mathbb{R}$). Definimos

$$\begin{aligned} B_\epsilon(x) &= \{y \in M \mid d(x, y) < \epsilon\} \quad \text{(Bola abierta de centro } x \text{ y radio } \epsilon). \\ B'_\epsilon(x) &= \{y \in M \mid d(x, y) \leq \epsilon\} \quad \text{(Bola cerrada de centro } x \text{ y radio } \epsilon). \end{aligned}$$

La figura muestra las bolas de centro $(0,0)$ y radio 1 para las tres métricas que hemos definido en $\mathbb{R}^2$.



Las bolas con otros centros son trasladadas de éstas, y las bolas de otros radios son homotéticas. Las bolas abiertas se diferencian de las cerradas en que las primeras no contienen los puntos del borde. El interés de las bolas reside en que una bola de centro un punto P contiene todos los puntos de alrededor de P, por pequeño que sea su radio. Observar que el concepto de "puntos de alrededor" es un tanto escurridizo: Según lo que acabamos de decir, ningún punto en particular (distinto de P) está alrededor de P, pues siempre podemos tomar una bola suficientemente pequeña como para que deje fuera a dicho punto. Esto significa que no podemos dar sentido a la afirmación "Q es un punto de alrededor de P", pero lo importante es que sí tiene sentido decir "El conjunto A contiene a todos los puntos de alrededor de P". Esto sucede cuando A contiene una bola cualquiera de centro P, y entonces diremos que A es un entorno de P. Aunque el concepto de entorno podría tomarse como concepto topológico básico, lo cierto es que es más cómodo partir de un concepto "más regular": diremos que un conjunto es abierto si es un entorno de todos sus puntos. Los conjuntos abiertos tienen las propiedades que recoge la definición siguiente:

Definición 1.9 Una *topología* en un conjunto X es una familia $\mathcal{T}$ de subconjuntos de X a cuyos elementos llamaremos *abiertos,* tal que cumpla las propiedades siguientes:

a) $\varnothing$ y X son abiertos.

b) La unión de cualquier familia de abiertos es un abierto.

c) La intersección de dos abiertos es un abierto.

Un *espacio topológico* es un par $(X,\mathcal{T})$, donde X es un conjunto y $\mathcal{T}$ es una topología en X. En la práctica escribiremos simplemente X en lugar de $(X,\mathcal{T})$.

Sea M un espacio métrico. Diremos que un conjunto $G \subset M$ es *abierto* si para todo $x \in G$ existe un $\epsilon > 0$ tal que $B_\epsilon(x) \subset G$. Es inmediato comprobar que los conjuntos abiertos así definidos forman una topología en M, a la que llamaremos *topología inducida por la métrica.* En lo sucesivo consideraremos siempre a los espacios métricos como espacios topológicos con esta topología.

En el párrafo previo a la definición de topología hemos definido "abierto" como un conjunto que es entorno de todos sus puntos. Puesto que formalmente

hemos definido los espacios topológicos a partir del concepto de abierto, ahora hemos de definir el concepto de entorno.

Si X es un espacio topológico, $U \subset X$ y $x \in U$, diremos que U es un *entorno* de x si existe un abierto G tal que $x \in G \subset U$.

Es inmediato comprobar que, en un espacio métrico, U es entorno de x si y sólo si existe un $\epsilon > 0$ tal que $B_\epsilon(x) \subset U$, es decir, si y sólo si U contiene a todos los puntos de alrededor de x, tal y como habíamos afirmado.

Ejemplo El intervalo $I = [0, 1]$, visto como subconjunto de $\mathbb{R}$, es entorno de todos sus puntos excepto de sus extremos 0 y 1, pues si $0 < x < 1$ siempre podemos tomar $\epsilon = \min\{x, 1-x\}$ y entonces $B_\epsilon(x) =]x-\epsilon, x+\epsilon[\subset I$. En cambio, I no contiene todos los puntos de alrededor de 1, pues toda bola de centro 1 contiene puntos a la derecha de 1 y ninguno de ellos está en I. El caso del 0 es similar. En particular I no es abierto. ■

El teorema siguiente recoge las propiedades básicas de los entornos. La prueba es inmediata.

Teorema 1.10 *Sea X un espacio topológico, $x \in X$ y E_x la familia de todos los entornos de x.*

a) *Un conjunto $G \subset X$ es abierto si y sólo si es un entorno de todos sus puntos.*

b1) $X \in E_x$.

b2) *Si $U \in E_x$ y $U \subset V \subset X$ entonces $V \in E_x$.*

b3) *Si U, $V \in E_x$ entonces $U \cap V \in E_x$.*

Puesto que los abiertos pueden definirse a partir de los entornos, es obvio que si dos topologías sobre un mismo conjunto tienen los mismos entornos entonces son iguales. Las desigualdades del teorema 1.7 implican que una bola para una de las tres distancias definidas en el producto M contiene otra bola del mismo centro para cualquiera de las otras distancias. De aquí se sigue que un subconjunto de M es entorno de un punto para una distancia si y sólo si lo es para las demás, y de aquí a su vez que las tres distancias definen la misma topología en el producto. En particular, las tres distancias que tenemos definidas sobre $\mathbb{K}^n$ definen la misma topología, a la que llamaremos *topología usual* o *topología euclídea* en $\mathbb{K}^n$.

Ésta es una primera muestra del carácter auxiliar de las distancias en topología. Cuando queramos probar un resultado puramente topológico sobre $\mathbb{R}^n$ podremos apoyarnos en la distancia que resulte más conveniente, sin que ello suponga una pérdida de generalidad. La distancia d_2 es la distancia euclídea y por lo tanto la más natural desde un punto de vista geométrico, pero las distancias d_1 y d_∞ son formalmente más sencillas y a menudo resultan más adecuadas.

Ejemplo Es fácil definir una distancia en $\mathbb{K}^n$ que induzca una topología distinta de la usual. De hecho, si X es un conjunto cualquiera podemos considerar la distancia $d : X \times X \longrightarrow \mathbb{R}$ dada por

$$d(x,y) = \begin{cases} 1 & \text{si } x \neq y, \\ 0 & \text{si } x = y \end{cases}$$

Es fácil ver que efectivamente es una distancia y para todo punto x se cumple que $B_1(x) = \{x\}$, luego $\{x\}$ es un entorno de x, luego es un abierto y, como toda unión de abiertos es abierta, de hecho todo subconjunto de X es abierto. La métrica d recibe el nombre de *métrica discreta* y la topología que induce es la *topología discreta.* Un espacio topológico cuya topología sea la discreta es un *espacio discreto.*

En un espacio discreto un punto no tiene más punto a su alrededor que él mismo. Esta topología es la más adecuada para conjuntos como $\mathbb{N}$ o $\mathbb{Z}$, pues, efectivamente, un número entero no tiene alrededor a ningún otro. ■

Las bolas abiertas de un espacio métrico son abiertas. Esto es fácil de ver intuitivamente, pero el mero hecho de que las hayamos llamado así no justifica que lo sean:

Teorema 1.11 *Las bolas abiertas de un espacio métrico son conjuntos abiertos.*

DEMOSTRACIÓN: Sea $B_\epsilon(x)$ una bola abierta y sea $y \in B_\epsilon(x)$. Entonces $d(x,y) < \epsilon$. Sea $0 < \delta < \epsilon - d(x,y)$. Basta probar que $B_\delta(y) \subset B_\epsilon(x)$. Ahora bien, si $z \in B_\delta(y)$, entonces $d(z,x) \leq d(z,y) + d(y,x) < \delta + d(x,y) < \epsilon$, luego en efecto $z \in B_\epsilon(x)$. ■

1.2 Bases y subbases

Hemos visto que la topología en un espacio métrico se define a partir de las bolas abiertas. El concepto de "bola abierta" no tiene sentido en un espacio topológico arbitrario en el que no tengamos dada una distancia, sin embargo hay otras familias de conjuntos que pueden representar un papel similar.

Definición 1.12 Sea X un espacio topológico. Diremos que una familia $\mathcal{B}$ de abiertos de X (a los que llamaremos *abiertos básicos*) es una *base* de X si para todo abierto G de X y todo punto $x \in G$ existe un abierto $B \in \mathcal{B}$ tal que $x \in B \subset G$.

Si $x \in X$ diremos que una familia E de entornos (abiertos) de x (a los que llamaremos *entornos básicos* de x) es una *base de entornos* (abiertos) de x si todo entorno de x contiene un elemento de E.

En estos términos la propia definición de los abiertos métricos (junto con el hecho de que las bolas abiertas son realmente conjuntos abiertos) prueba que las bolas abiertas son una base de la topología métrica, y también es claro que

las bolas abiertas de centro un punto x forman una base de entornos abiertos de x. Pero estos conceptos son mucho más generales. Pensemos por ejemplo que otras bases de un espacio métrico son las bolas abiertas de radio menor que 1, las bolas abiertas de radio racional, etc. Cualquier base determina completamente la topología y en cada ocasión puede convenir trabajar con una base distinta.

Teorema 1.13 *Sea X un espacio topológico.*

a) Una familia de abiertos $\mathcal{B}$ es una base de X si y sólo si todo abierto de X es unión de abiertos de $\mathcal{B}$.

b) Si $\mathcal{B}$ es una base de X y $x \in X$ entonces $\mathcal{B}_x = \{B \in \mathcal{B} \mid x \in B\}$ es una base de entornos abiertos de X.

c) Si para cada punto $x \in X$ el conjunto E_x es una base de entornos abiertos de x entonces $\mathcal{B} = \bigcup_{x \in X} E_x$ es una base de X.

DEMOSTRACIÓN: a) Si $\mathcal{B}$ es una base de X y G es un abierto es claro que G es la unión de todos los abiertos de $\mathcal{B}$ contenidos en G, pues una inclusión es obvia y si $x \in G$ existe un $B \in \mathcal{B}$ tal que $x \in B \subset G$, luego x está en la unión considerada. El recíproco es obvio.

b) Los elementos de $\mathcal{B}_x$ son obviamente entornos de x y si U es un entorno de x entonces existe un abierto G tal que $x \in G \subset$U, y a su vez existe $B \in \mathcal{B}$ tal que $x \in B \subset G$, luego $B \in \mathcal{B}_x$ y $B \subset U$. Esto prueba que $\mathcal{B}_x$ es una base de entornos abiertos de x.

c) Si G es un abierto de X y $x \in G$, entonces G es un entorno de x, luego existe un entorno básico $B \in E_x$ tal que $x \in B \subset G$ y ciertamente $B \in \mathcal{B}$. Como además los elementos de $\mathcal{B}$ son abiertos, tenemos que $\mathcal{B}$ es una base de X. ■

Una forma habitual de definir una topología en un conjunto es especificar una base o una base de entornos abiertos de cada punto. Por ejemplo, la topología métrica puede definirse como la topología que tiene por base a las bolas abiertas o como base de entornos de cada punto x a las bolas abiertas de centro x. No obstante, para que una familia de conjuntos pueda ser base de una topología ha de cumplir unas propiedades muy simples que es necesario comprobar. El teorema siguiente da cuenta de ellas.

Teorema 1.14 *Sea X un conjunto y $\mathcal{B}$ una familia de subconjuntos de X que cumpla las propiedades siguientes:*

a) $X = \bigcup_{B \in \mathcal{B}} B$,

b) Si $U, V \in \mathcal{B}$ y $x \in U \cap V$ entonces existe $W \in \mathcal{B}$ tal que $x \in W \subset U \cap V$.

Entonces existe una única topología en X para la cual $\mathcal{B}$ es una base.

DEMOSTRACIÓN: Definimos los abiertos de X como las uniones de elementos de $\mathcal{B}$. Basta comprobar que estos abiertos forman realmente una topología, pues ciertamente en tal caso $\mathcal{B}$ será una base y la topología será única.

El conjunto vacío es abierto trivialmente (o si se prefiere, por definición). El conjunto X es abierto por la propiedad a).

La unión de abiertos es obviamente abierta (una unión de uniones de elementos de $\mathcal{B}$ es al fin y al cabo una unión de elementos de $\mathcal{B}$).

Sean G_1 y G_2 abiertos y supongamos que $x \in G_1 \cap G_2$. Como G_1 es unión de elementos de $\mathcal{B}$ existe un $U \in \mathcal{B}$ tal que $x \in U \subset G_1$. Similarmente $x \in V \subset G_2$ con $V \in \mathcal{B}$. Por la propiedad b) existe $W \in \mathcal{B}$ tal que $x \in W \subset U \cap V \subset G_1 \cap G_2$. Así pues, x está en la unión de los conjuntos $W \in \mathcal{B}$ tales que $W \subset G_1 \cap G_2$, y la otra inclusión es obvia, luego $G_1 \cap G_2$ es unión de elementos de $\mathcal{B}$. ■

El teorema siguiente nos da las condiciones que hemos de comprobar para definir una topología a partir de una familia de bases de entornos abiertos.

Teorema 1.15 *Sea X un conjunto y para cada $x \in X$ sea $\mathcal{B}_x$ una familia no vacía de subconjuntos de X tal que:*

a) Si $U \in \mathcal{B}_x$, entonces $x \in U$.

b) Si $U, V \in \mathcal{B}_x$, existe un $W \in \mathcal{B}_x$ tal que $W \subset U \cap V$.

c) Si $x \in U \in \mathcal{B}_y$, existe un $V \in \mathcal{B}_x$ tal que $V \subset U$.

Entonces existe una única topología para la cual cada $\mathcal{B}_x$ es una base de entornos abiertos de x.

DEMOSTRACIÓN: Sea $\mathcal{B} = \bigcup\limits_{x \in X} \mathcal{B}_x$. Veamos que $\mathcal{B}$ cumple las condiciones del teorema anterior para ser base de una topología en X. Por la condición a) tenemos que $X = \bigcup\limits_{B \in \mathcal{B}} B$.

Si $U, V \in \mathcal{B}$ y $x \in U \cap V$, entonces $U \in \mathcal{B}_y$ y $V \in \mathcal{B}_z$ para ciertos y, z. Existen $U', V' \in \mathcal{B}_x$ tales que $U' \subset U$ y $V' \subset V$ (por la condición c). Existe $W \in \mathcal{B}_x$ tal que $W \subset U' \cap V'$ (por la condición b). Así $x \in W \subset U \cap V$ con $W \in \mathcal{B}$.

Por lo tanto $\mathcal{B}$ es la base de una topología en X para la que los elementos de cada $\mathcal{B}_x$ son abiertos y, en particular, entornos de x. Si A es un entorno de x para dicha topología, existe un $U \in \mathcal{B}$ tal que $x \in U \subset A$. Por definición de $\mathcal{B}$, existe un $y \in X$ tal que $U \in \mathcal{B}_y$, y por c) existe un $V \in \mathcal{B}_x$ tal que $V \subset U \subset A$. Esto prueba que $\mathcal{B}_x$ es una base de entornos de x.

Las bases de entornos determinan los entornos y por tanto la topología, es decir, se da la unicidad. ■

Ejemplo Como aplicación de este teorema vamos a convertir en espacio topológico al conjunto $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$. Para ello definimos como base de entornos abiertos de cada número real x al conjunto los entornos abiertos de x en $\mathbb{R}$ con la topología usual, la base de entornos abiertos de $+\infty$ está formada por los intervalos $]x, +\infty]$, donde x varía en $\mathbb{R}$, y la base de entornos abiertos de $-\infty$ la forman los intervalos $[-\infty, x[$, donde x varía en $\mathbb{R}$. Con esto estamos

diciendo que un conjunto contiene a los alrededores de $+\infty$ si contiene a todos los números reales a partir de uno dado, y análogamente para $-\infty$.

Es fácil comprobar que las familias consideradas cumplen las propiedades del teorema anterior, luego definen una topología en $\overline{\mathbb{R}}$.

Teniendo en cuenta que hemos definido los entornos abiertos de los números reales como los entornos abiertos que ya tienen en la topología usual, es inmediato que un subconjunto de $\mathbb{R}$ es abierto en la topología usual de $\mathbb{R}$ si y sólo si lo es en la topología que hemos definido en $\overline{\mathbb{R}}$. ■

Hay un concepto análogo a los de base y base de entornos que es menos intuitivo, pero mucho más práctico a la hora de definir topologías. Se trata del concepto de subbase:

Definición 1.16 Sea X un espacio topológico. Una familia de abiertos S es una *subbase* de X si las intersecciones finitas de elementos de S forman una base de X.

Por ejemplo, es fácil ver que los intervalos abiertos $]a, b[$ forman una base de $\mathbb{R}$ (son las bolas abiertas). Por consiguiente, los intervalos de la forma $]-\infty, a[$ y $]a, +\infty[$ forman una subbase de $\mathbb{R}$, pues son abiertos y entre sus intersecciones finitas se encuentran todos los intervalos $]a, b[$ (notar además que cualquier familia de abiertos que contenga a una base es una base).

La ventaja de las subbases consiste en que una familia no ha de cumplir ninguna propiedad en especial para ser subbase de una topología:

Teorema 1.17 *Sea X un conjunto y S una familia de subconjuntos de X. Entonces existe una única topología en X de la cual S es subbase.*

Demostración: Sea $\mathcal{B}$ la familia de las intersecciones finitas de elementos de S. Entonces $X = \bigcap\limits_{G \in \varnothing} G \in \mathcal{B}$ y obviamente la intersección de dos intersecciones finitas de elementos de S es una intersección finita de elementos de S; luego si U, $V \in \mathcal{B}$ también $U \cap V \in \mathcal{B}$, de donde se sigue que $\mathcal{B}$ es la base de una topología en X, de la cual S es subbase. Claramente es única, pues $\mathcal{B}$ es base de cualquier topología de la que S sea subbase. ■

1.3 Productos y subespacios

Hemos visto que el producto de una familia finita de espacios métricos es de nuevo un espacio métrico de forma natural (o mejor dicho, de tres formas distintas pero equivalentes desde un punto de vista topológico). Ahora veremos que la topología del producto se puede definir directamente a partir de las topologías de los factores sin necesidad de considerar las distancias. Más aún, podemos definir el producto de cualquier familia de espacios topológicos, no necesariamente finita.

Definición 1.18 Sean $\{X_i\}_{i\in I}$ espacios topológicos. Consideremos su producto cartesiano $X = \prod_{i\in I} X_i$ y las proyecciones $p_i : X \longrightarrow X_i$ que asignan a cada punto su coordenada i-ésima. Llamaremos *topología producto* en X a la que tiene por subbase a los conjuntos $p_i^{-1}[G]$, donde $i \in I$ y G es abierto en X_i.

Una base de la topología producto la forman los conjuntos de la forma $\bigcap_{i\in F} p_i^{-1}[G_i]$, donde F es un subconjunto finito de I y G_i es abierto en X_i.

Equivalentemente, la base está formada por los conjuntos $\prod_{i\in I} G_i$, donde cada G_i es abierto en X_i y $G_i = X_i$ salvo para un número finito de índices. Al conjunto de estos índices se le llama *soporte* del abierto básico $\prod_{i\in I} G_i$.

Si el número de factores es finito la restricción se vuelve vacía, de modo que un abierto básico en un producto $X_1 \times \cdots \times X_n$ es simplemente un conjunto de la forma $G_1 \times \cdots \times G_n$, donde cada G_i es abierto en X_i.

En lo sucesivo "casi todo i" querrá decir "todo índice i salvo un número finito de ellos".

Teorema 1.19 *Sean $\{X_i\}_{i\in I}$ espacios topológicos, para cada i sea $\mathcal{B}_i$ una base de X_i. Entonces los conjuntos de la forma $\prod_{i\in I} G_i$, donde cada G_i está en $\mathcal{B}_i$ o es X_i (y casi todos son X_i) forman una base de $\prod_{i\in I} X_i$.*

Demostración: Consideremos la topología $\mathcal{T}$ en el producto que tiene por subbase a los conjuntos $p_i^{-1}[G_i]$ con G_i en $\mathcal{B}_i$ (y, por consiguiente, tieen por base a los abiertos del enunciado). Como ciertamente estos conjuntos son abiertos para la topología producto, tenemos que todo abierto de $\mathcal{T}$ lo es de la topología producto. Recíprocamente, un abierto subbásico de la topología producto es $p_i^{-1}[G_i]$, con G_i abierto en X_i. Entonces $G_i = \bigcup_{B\in A_i} B$, donde cada A_i es un subconjunto de $\mathcal{B}_i$. Por lo tanto $p_i^{-1}[G_i] = \bigcup_{B\in A_i} p_i^{-1}[B]$ es abierto de $\mathcal{T}$. Por consiguiente todo abierto de la topología producto lo es de $\mathcal{T}$ y así ambas topologías coinciden. ∎

En lo sucesivo, a pesar de que en el producto se puedan considerar otras bases, cuando digamos "abiertos básicos" nos referiremos a los abiertos indicados en el teorema anterior tomando como bases de los factores las propias topologías salvo que se esté considerando alguna base en concreto.

Tal y como anunciábamos, el producto de espacios topológicos generaliza al producto de espacios métricos (o de espacios normados). El teorema siguiente lo prueba.

Teorema 1.20 *Si $M_1, \ldots, M_n$ son espacios métricos, entonces la topología inducida por las métricas de* 1.7 *en $M = M_1 \times \cdots \times M_n$ es la topología producto.*

Demostración: Como las tres métricas inducen la misma topología sólo es necesario considerar una de ellas, pero para la métrica d_∞ se cumple $B_\epsilon(x) =$

$B_\epsilon(x_1)\times\cdots\times B_\epsilon(x_n)$, luego la base inducida por la métrica es base de la topología producto. ■

La definición de topología producto es sin duda razonable para un número finito de factores. Sin embargo cuando tenemos infinitos factores hemos exigido una condición de finitud que no hemos justificado. En principio podríamos considerar en $\prod\limits_{i\in I} X_i$ la topología que tiene por base a los productos $\prod\limits_{i\in I} G_i$ con G_i abierto en X_i (sin ninguna restricción de finitud). Ciertamente estos conjuntos son base de una topología a la que se le llama *topología de cajas*, y el teorema siguiente muestra que no coincide con la topología producto que hemos definido. La topología producto resulta ser mucho más útil que la topología de cajas.

Teorema 1.21 *Sea $\{X_i\}_{i\in I}$ una familia de espacios topológicos. Los únicos abiertos en $\prod\limits_{i\in I} X_i$ de la forma $\prod\limits_{i\in I} G_i \neq \varnothing$ son los abiertos básicos, es decir, los que además cumplen que cada G_i es abierto y $G_i = X_i$ para casi todo i.*

Demostración: Supongamos que $\prod\limits_{i\in I} G_i$ es un abierto no vacío. Consideremos un punto $x \in \prod\limits_{i\in I} G_i$. Existirá un abierto básico $\prod\limits_{i\in I} H_i$ tal que $x \in \prod\limits_{i\in I} H_i \subset \prod\limits_{i\in I} G_i$, luego para cada índice i se cumplirá $x_i \in H_i \subset G_i$, y como casi todo H_i es igual a X_i, tenemos que $G_i = X_i$ para casi todo i. Además tenemos que G_i es un entorno de x_i, pero dado cualquier elemento $a \in G_i$ siempre podemos formar un $x \in \prod\limits_{i\in I} G_i$ tal que $x_i = a$, luego en realidad tenemos que G_i es un entorno de todos sus puntos, o sea, es abierto. ■

Nos ocupamos ahora de los subespacios de un espacio topológico. Es evidente que todo subconjunto N de un espacio métrico M es también un espacio métrico con la misma distancia restringida a $N\times N$. Por lo tanto tenemos una topología en M y otra en N. Vamos a ver que podemos obtener la topología de N directamente a partir de la de M, sin pasar por la métrica.

Teorema 1.22 *Sea X un espacio topológico (con topología $\mathcal{T}$) y $A \subset X$. Definimos $\mathcal{T}_A = \{G\cap A \mid G \in \mathcal{T}\}$. Entonces $\mathcal{T}_A$ es una topología en A llamada* topología relativa *a X (o topología inducida por X) en A. En lo sucesivo sobreentenderemos siempre que la topología de un subconjunto de un espacio X es la topología relativa.*

Demostración: $A = X\cap A \in \mathcal{T}_A$, $\varnothing = \varnothing \cap A \in \mathcal{T}_A$.

Sea $C \subset \mathcal{T}_A$. Para cada $G \in C$ sea $U_G = \{U \in \mathcal{T} \mid U\cap A = G\} \neq \varnothing$ y sea V_G la unión de todos los abiertos de U_G.

De este modo V_G es un abierto en X y $V_G \cap A = G$.

$$\bigcup_{G\in C} G = \bigcup_{G\in C} V_G \cap A, \quad \text{y} \quad \bigcup_{G\in C} V_G \in \mathcal{T}, \quad \text{luego } G \in \mathcal{T}_A.$$

Si U, $V \in \mathcal{T}_A$, $U = U'\cap A$ y $V = V'\cap A$ con U', $V' \in \mathcal{T}$. Entonces $U\cap V = U'\cap V'\cap A \in \mathcal{T}_A$, pues $U'\cap V' \in \mathcal{T}$. Así $\mathcal{T}_A$ es una topología en A. ■

Ejemplo Consideremos $I = [0, 1] \subset \mathbb{R}$. Resulta que $]1/2, 1]$ es abierto en I, pues $]1/2, 1] =]1/2, 2[\cap I$ y $]1/2, 2[$ es abierto en $\mathbb{R}$. Sin embargo $]1/2, 1]$ no es abierto en $\mathbb{R}$ porque no es entorno de 1. Intuitivamente, $]1/2, 1]$ no contiene a todos los puntos de alrededor de 1 en $\mathbb{R}$ (faltan los que están a la derecha de 1), pero sí contiene a todos los puntos de alrededor de 1 en I. ■

La relación entre espacios y subespacios viene perfilada por los teoremas siguientes. El primero garantiza que la topología relativa no depende del espacio desde el que relativicemos.

Teorema 1.23 *Si X es un espacio topológico (con topología $\mathcal{T}$) y $A \subset B \subset X$, entonces $\mathcal{T}_A = (\mathcal{T}_B)_A$.*

DEMOSTRACIÓN: Si U es abierto en $\mathcal{T}_A$, entonces $U = V \cap A$ con $V \in \mathcal{T}$, luego $V \cap B \in \mathcal{T}_B$ y $U = V \cap A = (V \cap B) \cap A \in (\mathcal{T}_B)_A$.

Si $U \in (\mathcal{T}_B)_A$, entonces $U = V \cap A$ con $V \in \mathcal{T}_B$, luego $V = W \cap B$ con $W \in \mathcal{T}$. Así pues, $U = W \cap B \cap A = W \cap A \in \mathcal{T}_A$. Por lo tanto $\mathcal{T}_A = (\mathcal{T}_B)_A$. ■

Teorema 1.24 *Si $\mathcal{B}$ es una base de un espacio X y $A \subset X$, entonces el conjunto $\{B \cap A \mid B \in \mathcal{B}\}$ es una base de A.*

DEMOSTRACIÓN: Sea U un abierto en A y $x \in U$. Existe un V abierto en X tal que $U = V \cap A$. Existe un $B \in \mathcal{B}$ tal que $x \in B \subset V$, luego $x \in B \cap A \subset V \cap A = U$. Por lo tanto la familia referida es base de A. ■

Similarmente se demuestra:

Teorema 1.25 *Si $\mathcal{B}_x$ es una base de entornos (abiertos) de un punto x de un espacio X y $x \in A \subset X$, entonces $\{B \cap A \mid B \in \mathcal{B}_x\}$ es una base de entornos (abiertos) de x en A.*

Teorema 1.26 *Sea M un espacio métrico y sea $A \subset M$. Entonces $d' = d|_{A \times A}$ es una distancia en A y la topología que induce es la topología relativa.*

DEMOSTRACIÓN: Una base para la topología inducida por la métrica de A sería la formada por las bolas

$$B_\epsilon^{d'}(x) = \{a \in A \mid d'(x, a) < \epsilon\} = \{a \in X \mid d(x, a) < \epsilon\} \cap A = B_\epsilon^d(x) \cap A,$$

pero éstas son una base para la topología relativa por el teorema 1.24. ■

Teorema 1.27 *Sea $\{X_i\}_{i \in I}$ una familia de espacios topológicos y para cada i sea $Y_i \subset X_i$. Entonces la topología inducida en $\prod\limits_{i \in I} Y_i$ por $\prod\limits_{i \in I} X_i$ es la misma que la topología producto de los $\{Y_i\}_{i \in I}$.*

(La base obtenida por el teorema 1.24 a partir de la base usual de la topología producto es claramente la base usual de la topología producto.)

Ejercicio: Probar que la topología que hemos definido en $\overline{\mathbb{R}}$ induce en $\mathbb{R}$ la topología euclídea.

1.4 Algunos conceptos topológicos

Dedicamos esta sección a desarrollar el lenguaje topológico, es decir, a introducir las características de un espacio y sus subconjuntos que pueden definirse a partir de su topología. Hasta ahora hemos visto únicamente los conceptos de abierto y entorno. Otro concepto importante es el dual conjuntista de "abierto":

Definición 1.28 Diremos que un subconjunto de un espacio topológico es *cerrado* si su complementario es abierto.

Por ejemplo, un semiplano (sin su recta frontera) es un conjunto abierto, mientras que un semiplano con su frontera es cerrado, pues su complementario es el semiplano opuesto sin su borde, luego es abierto. Pronto veremos que la diferencia entre los conjuntos abiertos y los cerrados es precisamente que los primeros no contienen a los puntos de su borde y los segundos contienen todos los puntos de su borde. En importante notar que un conjunto no tiene por qué ser ni abierto ni cerrado. Baste pensar en el intervalo $[0, 1[$.

Ejercicio: Sea $X = [0, 1] \cup]3, 4]$. Probar que $]3, 4]$ es a la vez abierto y cerrado en X.

Las propiedades de los cerrados se deducen inmediatamente de las de los abiertos:

Teorema 1.29 *Sea X un espacio topológico. Entonces:*

a) $\varnothing$ y X son cerrados.

b) La intersección de cualquier familia de cerrados es un cerrado.

c) La unión de dos cerrados es un cerrado.

Puesto que la unión de abiertos es abierta, al unir todos los abiertos contenidos en un conjunto dado obtenemos el mayor abierto contenido en él. Similarmente, al intersecar todos los cerrados que contienen a un conjunto dado obtenemos el menor cerrado que lo contiene:

Definición 1.30 Sea X un espacio topológico. Llamaremos *interior* de un conjunto $A \subset X$ al mayor abierto contenido en A. Lo representaremos por $\operatorname{int} A$ o $\mathring{A}$. Llamaremos *clausura* de A al menor cerrado que contiene a A. Lo representaremos por $\operatorname{cl} A$ o $\overline{A}$. Los puntos de $\mathring{A}$ se llaman *puntos interiores* de A, mientras que los de $\overline{A}$ se llaman *puntos adherentes* de A.

Así pues, para todo conjunto A tenemos que $\mathring{A} \subset A \subset \overline{A}$. El concepto de punto interior es claro: un punto x es interior a un conjunto A si y sólo si A es un entorno de x. Por ejemplo, en un semiplano cerrado, los puntos interiores son los que no están en el borde. El teorema siguiente nos caracteriza los puntos adherentes.

Teorema 1.31 *Sea X un espacio topológico y A un subconjunto de X. Un punto x es adherente a A si y sólo si todo entorno de x corta a A.*

DEMOSTRACIÓN: Supongamos que x es adherente a A. Sea U un entorno de x. Existe un abierto G tal que $x \in G \subset U$. Basta probar que $G \cap A \neq \varnothing$. Ahora bien, en caso contrario $X \setminus G$ sería un cerrado que contiene a A, luego $\overline{A} \subset X \setminus G$, mientras que $x \in \overline{A} \cap G$.

Recíprocamente, si x tiene esta propiedad entonces $x \in \overline{A}$, ya que de lo contrario $X \setminus \overline{A}$ sería un entorno de x que no corta a A. ■

Vemos, pues, que, como su nombre indica, los puntos adherentes a un conjunto A son los que "están pegados" a A, en el sentido de que tienen alrededor puntos de A. Por ejemplo, es fácil ver que los puntos adherentes a un semiplano abierto son sus propios puntos más los de su borde. Veamos ahora que el concepto de borde corresponde a una noción topológica general:

Definición 1.32 Sea X un espacio topológico y $A \subset X$. Llamaremos *frontera* de A al conjunto $\partial A = \overline{A} \cap \overline{X \setminus A}$.

Así, los puntos frontera de un conjunto son aquellos que tienen alrededor puntos que están en A y puntos que no están en A. Esto es claramente una definición general del "borde" de un conjunto. Por ejemplo, la frontera de un triángulo la forman los puntos de sus lados.

Teorema 1.33 *Sea X un espacio topológico. Se cumple:*

a) Si $A \subset X$ entonces $\mathring{A} \subset A \subset \overline{A}$, además $\mathring{A}$ es abierto y $\overline{A}$ es cerrado.

b) Si $A \subset B \subset X$ y A es abierto entonces $A \subset \mathring{B}$.

c) Si $A \subset B \subset X$ y B es cerrado entonces $\overline{A} \subset B$.

d) Si $A \subset B \subset X$ entonces $\mathring{A} \subset \mathring{B}$ y $\overline{A} \subset \overline{B}$.

e) Si $A, B \subset X$, entonces $\operatorname{int}(A \cap B) = \operatorname{int} A \cap \operatorname{int} B$, $\overline{A \cup B} = \overline{A} \cup \overline{B}$.

f) $A \subset X$ es abierto si y sólo si $A = \mathring{A}$, y es cerrado si y sólo si $A = \overline{A}$.

g) Si $A \subset B \subset X$, entonces $\overline{A}^B = \overline{A}^X \cap B$.

h) Si $A \subset X$, entonces $\operatorname{int}(X \setminus A) = X \setminus \operatorname{cl} A$ y $\operatorname{cl}(X \setminus A) = X \setminus \operatorname{int} A$.

DEMOSTRACIÓN: Muchas de estas propiedades son inmediatas. Probaremos sólo algunas.

e) Claramente $A \cup B \subset \overline{A} \cup \overline{B}$, y el segundo conjunto es cerrado, luego $\overline{A \cup B} \subset \overline{A} \cup \overline{B}$. Por otra parte es claro que $\overline{A} \subset \overline{A \cup B}$ y $\overline{B} \subset \overline{A \cup B}$, luego tenemos la otra inclusión. La prueba con interiores es idéntica.

g) Observemos en primer lugar que los cerrados de B son exactamente las intersecciones con B de los cerrados de X. En efecto, si C es cerrado en X entonces $X \setminus C$ es abierto en X, luego $B \cap (X \setminus C) = B \setminus C$ es abierto en B, luego $B \setminus (B \setminus C) = B \cap C$ es cerrado en B. El recíproco es similar.

Por definición, $\overline{A}^X$ es la intersección de todos los cerrados en X que contienen a A, luego $\overline{A}^X \cap B$ es la intersección de todas las intersecciones con B de los cerrados en X que contienen a A, pero estos son precisamente los cerrados de B que contienen a A, o sea, $\overline{A}^X \cap B$ es exactamente $\overline{A}^B$.

h) Tenemos que $A \subset \operatorname{cl} A$, luego $X \setminus \operatorname{cl} A \subset X \setminus A$ y el primero es abierto, luego $X \setminus \operatorname{cl} A \subset \operatorname{int}(X \setminus A)$.

Por otra parte $\operatorname{int}(X \setminus A) \subset X \setminus A$, luego $A \subset X \setminus \operatorname{int}(X \setminus A)$, y éste es cerrado, luego $\overline{A} \subset X \setminus \operatorname{int}(X \setminus A)$ y $\operatorname{int}(X \setminus A) \subset X \setminus \overline{A}$. ■

En la prueba de la propiedad g) hemos visto lo siguiente:

Teorema 1.34 *Si X es un espacio topológico y $A \subset X$, los cerrados en la topología relativa de A son las intersecciones con A de los cerrados de X.*

Conviene observar que el análogo a g) para interiores es falso. Es decir, no se cumple en general que $\mathring{A}^B = \mathring{A} \cap B$. Por ejemplo, es fácil ver que en $\mathbb{R}$ se cumple $\mathring{\mathbb{N}} = \varnothing$, luego $\mathring{\mathbb{N}} \cap \mathbb{Z} = \varnothing$, mientras que $\mathring{\mathbb{N}}^{\mathbb{Z}} = \mathbb{N}$.

Vamos a refinar el concepto de punto adherente. Hemos visto que los puntos adherentes a un conjunto A son aquellos que tienen alrededor puntos de A. Sucede entonces que todo punto $x \in A$ es trivialmente adherente, porque x es un punto de alrededor de x y está en A. Cuando eliminamos esta posibilidad trivial tenemos el concepto de punto de acumulación:

Definición 1.35 Sea X un espacio topológico y $A \subset X$. Diremos que un punto $x \in X$ es un *punto de acumulación* de A si todo entorno U de x cumple $(U \setminus \{x\}) \cap A \neq \varnothing$. El conjunto de puntos de acumulación de A se llama *conjunto derivado* de A y se representa por A'.

Ejemplo Consideremos el conjunto $A = \big\{1/n \mid n \in \mathbb{N} \setminus \{0\}\big\} \subset \mathbb{R}$. Es fácil ver que $\overline{A} = A \cup \{0\}$. Sin embargo, $A' = \{0\}$. En efecto, en general se cumple que $A' \subset \overline{A}$, pero ningún punto $1/n \in A$ es de acumulación, pues

$$\left]\frac{1}{n} - \frac{1}{n+1}, \frac{1}{n} + \frac{1}{n+1}\right[$$

es un entorno de $1/n$ que no corta a A salvo en este mismo punto. ■

Como ya hemos dicho, siempre es cierto que $A' \subset \overline{A}$. También es claro que un punto adherente que no esté en A ha de ser un punto de acumulación de A. En otras palabras, $\overline{A} = A \cup A'$. Los puntos de A pueden ser de acumulación o no serlo. Por ejemplo, todos los puntos de $[0, 1]$ son de acumulación, mientras que los puntos del ejemplo anterior no lo eran.

Definición 1.36 Sea X un espacio topológico y $A \subset X$. Los puntos de $A \setminus A'$ se llaman *puntos aislados* de A.

Un punto $x \in A$ es aislado si y sólo si tiene un entorno U tal que $U \cap A = \{x\}$. El entorno lo podemos tomar abierto, y entonces vemos que los puntos aislados de A son los puntos que son abiertos en la topología relativa. Vemos, pues, que un espacio es discreto si y sólo si todos sus puntos son aislados. Es el caso del ejemplo anterior.

Definición 1.37 Un subconjunto A de un espacio topológico X es *denso* si $\overline{A} = X$.

Aplicando la propiedad h) de 1.33 vemos que A es denso en X si y sólo si $X \setminus A$ tiene interior vacío, es decir, si y sólo si todo abierto de X corta a A. Esto significa que los puntos de A están "en todas partes". Por ejemplo, puesto que todo intervalo de números reales contiene números racionales e irracionales, es claro que $\mathbb{Q}$ y $\mathbb{R} \setminus \mathbb{Q}$ son densos en $\mathbb{R}$. De aquí se sigue fácilmente que $\mathbb{Q}^n$ y $(\mathbb{R} \setminus \mathbb{Q})^n$ son densos en $\mathbb{R}^n$.

Ejercicio: Probar que si A es abierto en un espacio X y D es denso en X entonces $A \cap D$ es denso en A.

Hay una propiedad que no cumplen todos los espacios topológicos, pero sí la práctica totalidad de espacios de interés.

Definición 1.38 Diremos que un espacio topológico X es un *espacio de Hausdorff* si para todo par de puntos distintos x, $y \in X$ existen abiertos disjuntos U y V tales que $x \in U$, $y \in V$ (se dice que los abiertos U y V *separan* a x e y).

Por ejemplo, si en un conjunto X con más de un punto consideramos la topología formada únicamente por los abiertos $\varnothing$ y X (topología trivial) obtenemos un espacio que no es de Hausdorff. Se trata de un espacio patológico donde todo punto está alrededor de cualquier otro. Aunque la topología trivial es ciertamente la más patológica posible, lo cierto es que todas las topologías no de Hausdorff comparten con ella su patología, y rara vez resultan de interés. Veamos las propiedades de los espacios de Hausdorff:

Teorema 1.39 *Se cumplen las propiedades siguientes:*

a) En un espacio de Hausdorff, todo conjunto finito es cerrado.

b) Todo espacio de Hausdorff finito es discreto.

c) Todo subespacio de un espacio de Hausdorff es un espacio de Hausdorff.

d) El producto de una familia de espacios de Hausdorff es un espacio de Hausdorff.

e) Todo espacio métrico es un espacio de Hausdorff.

f) Un espacio X es de Hausdorff si y sólo si la diagonal $\Delta = \{(x, x) \mid x \in X\}$ es cerrada en $X \times X$.

Demostración: a) Basta probar que todo punto $\{x\}$ en un espacio de Hausdorff X es cerrado. Ahora bien, dado $y \in X \setminus \{x\}$, existen abiertos disjuntos U, V tales que $x \in U$, $y \in V$, luego $y \in V \subset X \setminus \{x\}$, lo que prueba que $X \setminus \{x\}$ es entorno de todos sus puntos, luego $\{x\}$ es cerrado.

b) En un espacio de Hausdorff finito todo subconjunto es cerrado, luego todo subconjunto es abierto, luego es discreto.

c) Si X es un espacio de Hausdorff y $A \subset X$, dados dos puntos x, $y \in A$, existen abiertos disjuntos U, V en X que separan a x e y, luego $U \cap A$, $V \cap A$ son abiertos disjuntos en A que separan a x e y.

d) Consideremos un producto de espacios de Hausdorff $\prod\limits_{i \in I} X_i$ y dos de sus puntos x, y. Sea i_0 un índice tal que $x_{i_0} \neq y_{i_0}$. Existen abiertos U, V en X_{i_0} que separan a x_{i_0} e y_{i_0}. Entonces $p_{i_0}^{-1}(U)$ y $p_{i_0}^{-1}(V)$ son abiertos subbásicos disjuntos en el producto que separan a x e y.

e) Si X es un espacio métrico, dos de sus puntos x, y están separados por las bolas de centros x, y y radio $d(x, y)/2$.

f) La diagonal Δ es cerrada si y sólo si su complementario es abierto, si y sólo si para todo par $(x, y) \in X \times X$ con $x \neq y$ existe un abierto básico $U \times V$ en $X \times X$ tal que $(x, y) \in U \times V \subset X \times X \setminus \Delta$. Ahora bien, la condición $U \times V \subset X \times X \setminus \Delta$ equivale a $U \cap V = \varnothing$, luego la diagonal es cerrada si y sólo si X es Hausdorff. ■

Ejercicio: Probar que si un producto de espacios topológicos es un espacio de Hausdorff no vacío, entonces cada uno de los factores es un espacio de Hausdorff.

Terminamos la sección con algunas propiedades métricas, no topológicas, es decir propiedades definidas a partir de la distancia en un espacio métrico y que no se pueden expresar en términos de su topología.

Definición 1.40 Un subconjunto A de un espacio métrico es *acotado* si existe un $M > 0$ tal que para todo par de puntos x, $y \in A$ se cumple $d(x, y) \leq M$. El *diámetro* de un conjunto acotado A es el supremo de las distancias $d(x, y)$ cuando (x, y) varía en $A \times A$.

Es fácil probar que todo subconjunto de un conjunto acotado es acotado, así como que la unión finita de conjuntos acotados está acotada. Sin embargo hemos de tener presente el hecho siguiente: dado un espacio métrico M, podemos definir $d'(x, y) = \text{mín}\big\{1, d(x, y)\big\}$. Es fácil ver que d' es una distancia en M y las bolas de radio menor que 1 para d' coinciden con las bolas respecto a d. Como estas bolas forman una base de las respectivas topologías métricas concluimos que ambas distancias definen la misma topología. Sin embargo, respecto a d' todos los conjuntos están acotados. Esto prueba que el concepto de acotación no es topológico.

Ejercicio: Calcular el diámetro de una bola abierta en $\mathbb{R}^n$ y en un espacio con la métrica discreta.

Definición 1.41 Si M es un espacio métrico, $A \neq \varnothing$ un subconjunto de M y $x \in M$, definimos la distancia de x a A como $d(x,A) = \text{ínf}\{d(x,y) \mid y \in A\}$.

Es evidente que si $x \in A$ entonces $d(x,A) = 0$, pues entre las distancias cuyo ínfimo determinan $d(x,A)$ se encuentra $d(x,x) = 0$. Sin embargo los puntos que cumplen $d(x,A) = 0$ no están necesariamente en A.

Teorema 1.42 *Si M es un espacio métrico y $A \subset M$, entonces un punto x cumple $d(x,A) = 0$ si y sólo si x es adherente a A.*

DEMOSTRACIÓN: Si $d(x,A) = 0$, para probar que es adherente basta ver que toda bola abierta de centro x corta a A. Dado $\epsilon > 0$ tenemos que $d(x,A) < \epsilon$, lo que significa que existe un $y \in A$ tal que $d(x,y) < \epsilon$, es decir, $y \in B_\epsilon(x) \cap A$. El recíproco se prueba igualmente. ■

En el caso de espacios normados podemos hacer algunas afirmaciones adicionales. La prueba del teorema siguiente es inmediata.

Teorema 1.43 *Sea E un espacio normado y $A \subset E$. Las afirmaciones siguientes son equivalentes:*

a) A es acotado.

b) Existe un $M > 0$ tal que $\|x\| \leq M$ para todo $x \in A$.

c) Existe un $M > 0$ tal que $A \subset B_M(0)$.

Ejercicio: Probar que en un espacio normado la clausura de una bola abierta es la bola cerrada del mismo radio y el interior de una bola cerrada es la bola abierta. ¿Cuál es la frontera de ambas? Dar ejemplos que muestren la falsedad de estos hechos en un espacio métrico arbitrario.

1.5 Continuidad

Finalmente estamos en condiciones de formalizar la idea de función continua como función f que envía los alrededores de un punto a los alrededores de su imagen. No exigimos que las imágenes de los puntos de alrededor de un punto x sean todos los puntos de alrededor de $f(x)$. Por ejemplo, sea S la circunferencia unidad en $\mathbb{R}^2$ y f la aplicación dada por

$$\begin{array}{ccc} [0,1] & \longrightarrow & S \\ x & \mapsto & (\cos 2\pi x, \operatorname{sen} 2\pi x) \end{array}$$

Lo que hace f es "pegar" los extremos del intervalo en un mismo punto $(1,0)$. Queremos que esta aplicación sea continua, y vemos que $[0,1/4]$ es un entorno de 0 que se transforma en "medio" entorno de $(1,0)$, en los puntos de alrededor de $(1,0)$ contenidos en el semiplano $y > 0$. Esto no debe ser, pues, un obstáculo a la continuidad.

Pedir que los puntos de alrededor de x sean enviados a puntos de alrededor de $f(x)$ (no necesariamente todos) es pedir que todo entorno U de $f(x)$ contenga a las imágenes de los puntos de alrededor de x, es decir, las imágenes de un entorno de x, es decir, que $f^{-1}[U]$ contenga un entorno de x, pero esto equivale a que él mismo lo sea. Así pues:

Definición 1.44 Una aplicación $f : X \longrightarrow Y$ entre dos espacios topológicos es *continua* en un punto $x \in X$ si para todo entorno U de $f(x)$ se cumple que $f^{-1}[U]$ es un entorno de x. Diremos que f es *continua* si lo es en todos los puntos de x.

Observar que en esta definición podemos sustituir "entorno" por "entorno básico". En un espacio métrico podemos considerar concretamente bolas abiertas, y entonces la definición se particulariza como sigue:

Teorema 1.45 *Una aplicación $f : M \longrightarrow N$ entre dos espacios métricos es continua en un punto $x \in M$ si y sólo si para todo $\epsilon > 0$ existe un $\delta > 0$ tal que si $d(x, x') < \delta$ entonces $d\big(f(x), f(x')\big) < \epsilon$.*

Veamos varias caracterizaciones de la continuidad.

Teorema 1.46 *Sea $f : X \longrightarrow Y$ una aplicación entre espacios topológicos. Las afirmaciones siguientes son equivalentes:*

a) f es continua.

b) Para todo abierto (básico) G de Y se cumple que $f^{-1}[G]$ es abierto en X.

c) Para todo cerrado C de Y se cumple que $f^{-1}[C]$ es cerrado en X.

d) Para todo $A \subset X$ se cumple $f[\overline{A}] \subset \overline{f[A]}$.

e) Para todo $B \subset Y$ se cumple $\overline{f^{-1}[B]} \subset f^{-1}[\overline{B}]$.

f) Para todo $B \subset Y$ se cumple $f^{-1}[\operatorname{int} B] \subset \operatorname{int} f^{-1}[B]$.

Demostración: a) $\leftrightarrow$ b). Si f es continua y $x \in f^{-1}[G]$ entonces $f(x) \in G$, luego G es un entorno de $f(x)$, luego por definición de continuidad $f^{-1}[G]$ es un entorno de x, luego $f^{-1}[G]$ es abierto. Es claro que b) $\rightarrow$ a).

Evidentemente b) $\leftrightarrow$ c).

a) $\rightarrow$ d). Si $x \in f[\overline{A}]$ entonces $x = f(y)$, con $y \in \overline{A}$. Si E es un entorno de x, por definición de continuidad $f^{-1}[E]$ es un entorno de y, luego $f^{-1}[E] \cap A \neq \varnothing$, de donde $E \cap f[A] \neq \varnothing$, lo que prueba que $x \in \overline{f[A]}$.

d) $\rightarrow$ e). Tenemos que $f^{-1}[B] \subset X$, luego $f\big[\overline{f^{-1}[B]}\big] \subset \overline{f\big[f^{-1}[B]\big]} \subset \overline{B}$, luego $\overline{f^{-1}[B]} \subset f^{-1}[\overline{B}]$.

e) $\rightarrow$ f). En efecto:

$$\begin{aligned} f^{-1}[\operatorname{int} B] &= f^{-1}\big[Y\setminus (Y\setminus \operatorname{int} B)\big] = X\setminus f^{-1}[Y\setminus \operatorname{int} B] \\ &= X\setminus f^{-1}[\overline{Y\setminus B}] \subset X\setminus \overline{f^{-1}[Y\setminus B]} = X\setminus \overline{X\setminus f^{-1}[B]} \\ &= X\setminus (X\setminus \operatorname{int} f^{-1}[B]) = \operatorname{int} f^{-1}[B]. \end{aligned}$$

f) $\rightarrow$ b). Si B es abierto en Y, entonces

$$f^{-1}[B] = f^{-1}[\operatorname{int} B] \subset \operatorname{int} f^{-1}[B] \subset f^{-1}[B],$$

luego $f^{-1}[B] = \operatorname{int} f^{-1}[B]$ que es, por lo tanto, abierto. ■

Ahora vamos a probar una serie de resultados generales que nos permitirán reconocer en muchos casos la continuidad de una aplicación de forma inmediata. De la propia definición de continuidad se sigue inmediatamente:

Teorema 1.47 *Si $f : X \longrightarrow Y$ es continua en un punto x y $g : Y \longrightarrow Z$ es continua en $f(x)$, entonces $f\circ g$ es continua en x. En particular, la composición de aplicaciones continuas es una aplicación continua.*

Otro hecho básico es que la continuidad depende sólo de la topología en la imagen y no de la del espacio de llegada.

Teorema 1.48 *Sea $f : X \longrightarrow Y$ una aplicación entre espacios topológicos. Entonces f es continua en un punto $x\in X$ como aplicación $f : X \longrightarrow Y$ si y sólo si lo es como aplicación $f : X \longrightarrow f[X]$.*

Demostración: Un entorno de $f(x)$ en $f[X]$ es $U\cap f[X]$, donde U es un entorno de $f(x)$ en Y, pero $f^{-1}[U\cap A] = f^{-1}[U]$, luego es indistinto considerar entornos en $f[X]$ o en Y. ■

Teniendo en cuenta que la aplicación identidad en un conjunto es obviamente continua, de los teoremas anteriores se deduce inmediatamente el que sigue:

Teorema 1.49 *Si X es un espacio topológico y $A\subset X$, entonces la inclusión $i : A \longrightarrow X$ dada por $i(x) = x$ es continua. Por tanto, si $f : X \longrightarrow Y$ es continua en un punto $x\in A$, la restricción $f|_A = i\circ f$ es continua en x.*

En particular la restricción de una aplicación continua a un subconjunto es también continua. El recíproco no es cierto, pero se cumple lo siguiente:

Teorema 1.50 *Dada una aplicación $f : X \longrightarrow Y$, si A es un entorno de un punto $x\in X$ y $f|_A$ es continua en x, entonces f es continua en x.*

Demostración: Si U es un entorno de $f(x)$ en Y, entonces $(f|_A)^{-1}[U] = f^{-1}[U]\cap A$ es un entorno de x en A, luego existe un entorno G de x en X de manera que $x\in G\cap A = f^{-1}[U]\cap A$, luego en particular $x\in G\cap A\subset f^{-1}[U]$, y $G\cap A$ es un entorno de x en X, luego $f^{-1}[U]$ también lo es. ■

Esto significa que la continuidad es una propiedad local, es decir, el que una función sea continua o no en un punto es un hecho que sólo depende del comportamiento de la función en un entorno del punto. En particular, si cubrimos un espacio topológico por una familia de abiertos, para probar que una aplicación es continua basta ver que lo es su restricción a cada uno de los abiertos. Esto es cierto también si cubrimos el espacio con cerrados a condición de que sean un número finito.

Teorema 1.51 *Sea $f : X \longrightarrow Y$ una aplicación entre espacios topológicos. Sean $C_1, \ldots, C_n$ subconjuntos cerrados de X tales que $X = C_1 \cup \cdots \cup C_n$. Entonces f es continua si y sólo si cada $f|_{C_i}$ es continua.*

DEMOSTRACIÓN: Una implicación es obvia. Si las restricciones son continuas, entonces dado un cerrado C de Y, se cumple que

$$f^{-1}[C] = (f^{-1}[C] \cap C_1) \cup \cdots \cup (f^{-1}[C] \cap C_n) = (f|_{C_1})^{-1}[C] \cup \cdots \cup (f|_{C_n})^{-1}[C].$$

Ahora, cada $(f|_{C_i})^{-1}[C]$ es cerrado en C_i, luego es la intersección con C_i de un cerrado de X, luego es la intersección de dos cerrados en X, luego es cerrado en X. Así pues, $f^{-1}[C]$ es la unión de un número finito de cerrados de X, luego es cerrado en X. Esto prueba que f es continua. ∎

Teorema 1.52 *Si $\{X_i\}_{i \in I}$ es una familia de espacios topológicos, las proyecciones $p_i : \prod\limits_{i \in I} X_i \longrightarrow X_i$ son funciones continuas.*

DEMOSTRACIÓN: Las antiimágenes de abiertos en X_i son abiertos básicos del producto. ∎

Ejercicio: Probar que la topología producto es la menor topología que hace continuas a las proyecciones.

Teorema 1.53 *Si $\{X_i\}_{i \in I}$ es una familia de espacios topológicos y X es un espacio topológico, entonces una aplicación $f : X \longrightarrow \prod\limits_{i \in I} X_i$ es continua si y sólo si lo son todas las funciones $f_i = f \circ p_i$.*

DEMOSTRACIÓN: Si f es continua las funciones $f \circ p_i$ también lo son por ser composición de funciones continuas.

Si cada $f \circ p_i$ es continua, sea $A = \prod\limits_{i \in I} A_i$ un abierto básico del producto.

Sean $i_1, \ldots, i_n$ los índices tales que $A_{i_j} \neq X_{i_j}$. Entonces $f^{-1}\big[p_{i_j}^{-1}[A_{i_j}]\big] = (f \circ p_{i_j})^{-1}[A_{i_j}]$ es abierto en X, pero

$$A = \bigcap_{j=1}^{n} p_{i_j}^{-1}[A_{i_j}] \quad \text{y} \quad f^{-1}[A] = \bigcap_{j=1}^{n} f^{-1}\big[p_{i_j}^{-1}[A_{i_j}]\big]$$

es abierto en X. ∎

Así, por ejemplo, para probar que la aplicación $f : \mathbb{R} \longrightarrow \mathbb{R}^2$ dada por $f(x) = (x+1, x^2)$ es continua, basta probar que lo son las aplicaciones $x+1$ y x^2.

Definición 1.54 Sean E y F espacios normados. Una aplicación $f : E \longrightarrow F$ tiene la propiedad de Lipschitz si existe un $M > 0$ tal que para todos los vectores v, $w \in E$ se cumple que $\|f(v) - f(w)\| \leq M\|v - w\|$.

Teorema 1.55 *Las aplicaciones con la propiedad de Lipschitz son continuas.*

Demostración: Sea $f : E \longrightarrow F$ una aplicación con la propiedad de Lipschitz con constante M. Vamos a aplicar el teorema 1.45. Dado $\epsilon > 0$ tomamos $\delta = \epsilon/M$. Así, si $\|v - w\| < \delta$, entonces $\|f(v) - f(w)\| \leq M\|v - w\| < \epsilon$.

■

Por ejemplo, es fácil ver que si E es un espacio normado entonces la norma $\| \ \| : E \longrightarrow \mathbb{R}$ tiene la propiedad de Lipschitz con constante $M = 1$, luego es una aplicación continua. Un ejemplo menos trivial es el de la suma:

Teorema 1.56 *Sea E un espacio normado. Entonces la suma $+ : E \times E \longrightarrow E$ tiene la propiedad de Lipschitz, luego es continua.*

Demostración: Consideraremos a $E \times E$ como espacio normado con la norma $\| \ \|_1$. Entonces, si (u, v), $(a, b) \in E \times E$, tenemos que

$$\begin{aligned} \|(u+v) - (a+b)\| &= \|(u-a) + (v-b)\| \leq \|u-a\| + \|v-b\| \\ &= \|(u-a, v-b)\|_1 = \|(u,v) - (a,b)\|_1. \end{aligned}$$

■

El producto no cumple la propiedad de Lipschitz, pero aun así es continuo.

Teorema 1.57 *Sea E un espacio normado. El producto $\cdot : \mathbb{K} \times E \longrightarrow E$ es una aplicación continua.*

Demostración: Veamos que el producto es continuo en un punto (λ, x) de $\mathbb{K} \times E$. Usaremos la norma $\| \ \|_\infty$ en $\mathbb{K} \times E$. Dado $\epsilon > 0$, sea $(\lambda', x') \in \mathbb{K} \times E$.

$$\|\lambda' x' - \lambda x\| = \|\lambda'(x' - x) + (\lambda' - \lambda)x\| \leq |\lambda'| \, \|x' - x\| + |\lambda' - \lambda| \, \|x\|.$$

Tomemos ahora $0 < \delta < 1$ que cumpla además

$$\delta < \frac{\epsilon}{|\lambda| + \|x\| + 1} < \epsilon.$$

Así si $\|(\lambda', x') - (\lambda, x)\|_\infty < \delta$, entonces $|\lambda' - \lambda| < \delta$ y $\|x' - x\| < \delta$. En particular $|\lambda'| \leq |\lambda| + \delta < |\lambda| + 1$. Así

$$\|\lambda' x' - \lambda x\| < \frac{|\lambda'|\epsilon}{|\lambda| + \|x\| + 1} + \frac{\|x\| \, \epsilon}{|\lambda| + \|x\| + 1} < \epsilon.$$

■

Definición 1.58 Un *espacio vectorial topológico* es un par $(V, \mathcal{T})$, donde V es un espacio vectorial sobre $\mathbb{K}$ y $\mathcal{T}$ es una topología de Hausdorff sobre V de modo que las aplicaciones $+ : V \times V \longrightarrow V$ y $\cdot : \mathbb{K} \times V \longrightarrow V$ son continuas.

Hemos demostrado que todo espacio normado es un espacio vectorial topológico.

En general, si X es un conjunto y V es un espacio vectorial sobre un cuerpo K, el conjunto V^X de todas las aplicaciones de X en V es un espacio vectorial con las operaciones dadas por $(f+g)(x) = f(x) + g(x)$ y $(\alpha f)(x) = \alpha f(x)$.

Si X e Y son espacios topológicos, llamaremos $C(X, Y)$ al conjunto de todas las aplicaciones continuas de X en Y.

Si X es un espacio topológico y V un espacio vectorial topológico, entonces $C(X, V)$ resulta ser un subespacio de V^X, pues la suma de dos funciones continuas f y g puede obtenerse como composición de la aplicación $h : X \longrightarrow V \times V$ dada por $h(x) = \big(f(x), g(x)\big)$, que es continua, y la suma en V, que también lo es, luego $f + g$ es continua.

Similarmente se prueba que si $\alpha \in \mathbb{K}$ y $f \in C(X, V)$, entonces $\alpha f \in C(X, V)$.

El espacio $\mathbb{K}^X$ tiene también estructura de anillo conmutativo y unitario con el producto dado por $(fg)(x) = f(x)g(x)$. Como el producto $\cdot : \mathbb{K} \times \mathbb{K} \longrightarrow \mathbb{K}$ es continuo (tomando $E = \mathbb{K}$ en el teorema anterior), resulta que $C(X, \mathbb{K})$ es un subanillo de $\mathbb{K}^X$.

En general, una cuádrupla $(A, +, \cdot, \circ)$, donde la terna $(A, +, \cdot)$ es un espacio vectorial sobre un cuerpo K, la terna $(A, +, \circ)$ es un anillo unitario y además $\alpha(ab) = (\alpha a)b = a(\alpha b)$ para todo $\alpha \in K$, y todos los a, $b \in A$, se llama una K-álgebra.

Tenemos, pues, que si X es un espacio topológico, entonces $C(X, \mathbb{K})$ es una $\mathbb{K}$-álgebra de funciones. El espacio $C(X, \mathbb{K})$ contiene un subcuerpo isomorfo a $\mathbb{K}$, a saber, el espacio de las funciones constantes. Cuando no haya confusión identificaremos las funciones constantes con los elementos de $\mathbb{K}$, de modo que 2 representará a la función que toma el valor 2 sobre todos los puntos de X.

Más aún, si $p(x_1, \ldots, x_n) \in \mathbb{K}[x_1, \ldots, x_n]$, el polinomio p determina una única función evaluación $p : \mathbb{K}^n \longrightarrow \mathbb{K}$, de modo que los polinomios constantes se corresponden con las funciones constantes y la indeterminada x_i con la proyección en la i-ésima coordenada. Como todo polinomio es combinación de sumas y productos de constantes e indeterminadas, tenemos que $\mathbb{K}[x_1, \ldots, x_n] \subset C(\mathbb{K}^n, \mathbb{K})$.

Por ejemplo, la aplicación $f : \mathbb{R}^3 \longrightarrow \mathbb{R}^2$ dada por $f(x, y, z) = (3x - 2yz, xyz)$ es claramente continua.

Teorema 1.59 *La aplicación $h : \mathbb{K} \setminus \{0\} \longrightarrow \mathbb{K}$ dada por $h(x) = 1/x$ es continua.*

Demostración: Sea $x \in \mathbb{K} \setminus \{0\}$. Sea $\delta = |x|/2$. Si $y \in B_\delta(x)$ entonces $|-x-(y-x)| \geq \big||-x| - |y-x|\big|$, o sea, $|y| \geq |x| - \delta = \delta$.

Dado $\epsilon > 0$, sea $\delta' < 1$, $\delta' < \delta|x|\epsilon$. Así, si $|y - x| < \delta'$ tenemos

$$\left|\frac{1}{y} - \frac{1}{x}\right| = \frac{|y-x|}{|y|\,|x|} < \frac{\delta|x|\epsilon}{\delta|x|} = \epsilon.$$

■

Consecuentemente, si $f \in C(X, \mathbb{K})$ y f no se anula en X, podemos definir la función $1/f \in C(X, \mathbb{K})$ mediante $(1/f)(x) = 1/f(x)$.

Por ejemplo, la función $f : \mathbb{R} \setminus \{1\} \longrightarrow \mathbb{R}$ dada por

$$f(x) = \frac{x^2}{x-1}$$

es obviamente continua.

Teorema 1.60 *La función* $\sqrt{\ }: [0, +\infty[\longrightarrow [0, +\infty[$ *es continua.*

Demostración: Basta notar que

$$\sqrt{x} \in \,]a, b[\leftrightarrow a < \sqrt{x} < b \leftrightarrow a^2 < x < b^2 \leftrightarrow x \in \left]a^2, b^2\right[.$$

Esto significa que la antiimagen del intervalo $]a, b[$ es el intervalo $\left]a^2, b^2\right[$. Similarmente se ve que la antiimagen de un intervalo $[0, b[$ es el intervalo $\left[0, b^2\right[$ y, como estos intervalos constituyen una base de $[0, +\infty[$, tenemos que $\sqrt{\ }$ es una función continua. ■

Teorema 1.61 *Sea M un espacio métrico y $A \neq \varnothing$ un subconjunto de M. Entonces las aplicaciones $d : M \times M \longrightarrow \mathbb{R}$ y $d(\ \ , A) : M \longrightarrow \mathbb{R}$ son continuas.*

Demostración: Consideremos en $M \times M$ la distancia d_∞. Dado un par $(x, y) \in M \times M$, y un $\epsilon > 0$, basta probar que si (x', y') dista de (x, y) menos de $\epsilon/2$, es decir, si se cumple $d(x, x') < \epsilon/2$ y $d(y, y') < \epsilon/2$, entonces tenemos $|d(x, y) - d(x', y')| < \epsilon$. En efecto, en tal caso

$$d(x, y) \leq d(x, x') + d(x', y') + d(y', y) < d(x', y') + \epsilon,$$

luego $d(x, y) - d(x', y') < \epsilon$, e igualmente se llega a $d(x', y') - d(x, y) < \epsilon$, luego efectivamente $|d(x, y) - d(x', y')| < \epsilon$.

Para probar la continuidad de $d(\ \ , A)$ en un punto x observamos que

$$|d(x, A) - d(y, A)| \leq d(x, y).$$

En efecto, para todo $z \in A$ se cumple $d(x, z) \leq d(x, y) + d(y, z)$. De aquí se sigue claramente $d(x, A) \leq d(x, y) + d(y, z)$, y tomando el ínfimo en z vemos que $d(x, A) \leq d(x, y) + d(y, A)$. Similarmente se prueba $d(y, A) \leq d(x, y) + d(x, A)$, de donde se sigue la relación con valores absolutos. A su vez esta relación implica que si $d(x, y) < \epsilon$, entonces $|d(x, A) - d(y, A)| < \epsilon$, lo que expresa la continuidad de la aplicación. ■

Para terminar con las propiedades generales de las aplicaciones continuas probaremos un hecho de interés teórico:

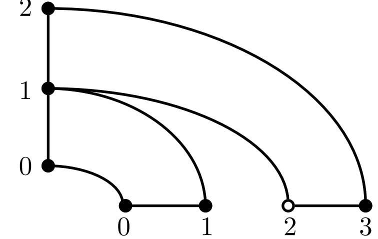
Teorema 1.62 *Sean* $f, g : X \longrightarrow Y$ *aplicaciones continuas,* $D \subset X$ *un conjunto denso tal que* $f|_D = g|_D$ *y supongamos que* Y *es un espacio de Hausdorff. Entonces* $f = g$.

Demostración: Sea $h : X \longrightarrow Y \times Y$ dada por $h(x) = \big(f(x), g(x)\big)$. Claramente es continua y $h[D] \subset \Delta$, donde $\Delta = \{(y, y) \mid y \in Y\}$ es un cerrado en $Y \times Y$ (por 1.39). Por lo tanto $h[X] = h[\overline{D}] \subset \overline{h[D]} \subset \Delta$, de donde $f = g$.
■

Definición 1.63 Una aplicación biyectiva $f : X \longrightarrow Y$ entre dos espacios topológicos es un *homeomorfismo* si f y su inversa son ambas continuas. Dos espacios topológicos son *homeomorfos* si existe un homeomorfismo entre ellos.

Un homeomorfismo induce una biyección entre los abiertos de los dos espacios, por lo que ambos son topológicamente indistinguibles. Es claro que cualquier propiedad definida exclusivamente a partir de la topología se conserva por homeomorfismos, luego dos espacios homeomorfos tienen las mismas propiedades topológicas.

Es importante notar que no toda biyección continua es un homeomorfismo. Por ejemplo, la identidad $I : \mathbb{R} \longrightarrow \mathbb{R}$, cuando en el primer espacio consideramos la topología discreta y en el segundo la euclídea es biyectiva y continua, pero no un homeomorfismo. Veamos un ejemplo más intuitivo:



Ejemplo Sea $f : [0, 1] \cup]2, 3] \longrightarrow [0, 2]$ la aplicación dada por

$$f(x) = \begin{cases} x & \text{si } 0 \leq x \leq 1 \\ x - 1 & \text{si } 2 < x \leq 3 \end{cases}$$

Es fácil ver que f es biyectiva, y es continua porque sus restricciones a los abiertos $[0, 1]$ y $]2, 3]$ de su dominio son ambas continuas (son polinomios). Sin embargo no es un homeomorfismo. La aplicación f^{-1} es continua en todos los puntos de $[0, 2]$ excepto en $x = 1$. En efecto, $[0, 1]$ es un entorno de $f^{-1}(1) = 1$, pero la antiimagen de este intervalo es el mismo $[0, 1]$, que no es un entorno de 1 en $[0, 2]$. En los demás puntos es continua, pues f^{-1} restringida a los abiertos de su dominio $[0, 1[$ y $]1, 2]$ es polinómica.

Lo que sucede es que, a pesar de ser biyectiva, f está "pegando" los intervalos $[0, 1]$ y $]2, 3]$ en el intervalo $[0, 2]$, por lo que f^{-1} "corta" éste por el punto 1. En general, si una aplicación continua es una aplicación que no corta, aunque puede pegar, un homeomorfismo es una aplicación que no corta ni pega. Para que no pegue ha de ser biyectiva, pero acabamos de ver que esto no es suficiente. El ejemplo de la topología discreta se interpreta igual: los puntos de $\mathbb{R}$ con la topología discreta están todos "separados" entre sí, luego al pasar a $\mathbb{R}$ con la topología euclídea los estamos "pegando", aunque no identifiquemos puntos.
■

Definición 1.64 Una aplicación $f : X \longrightarrow Y$ entre dos espacios topológicos es *abierta* si para todo abierto A de X, se cumple que $f[A]$ es abierto en Y.

De este modo, un homeomorfismo es una biyección continua y abierta. La propiedad de ser abierta no es muy intuitiva y tampoco es muy frecuente (salvo en el caso de los homeomorfismos). Sin embargo es de destacar el hecho siguiente:

Teorema 1.65 *Las proyecciones de un espacio producto en cada uno de sus factores son aplicaciones abiertas.*

Demostración: Basta ver que la proyección de un abierto básico es un abierto, pero los abiertos básicos son productos de abiertos, y las proyecciones son sus factores. ■

Definición 1.66 La *gráfica* de una aplicación $f : X \longrightarrow Y$ es el conjunto[1]

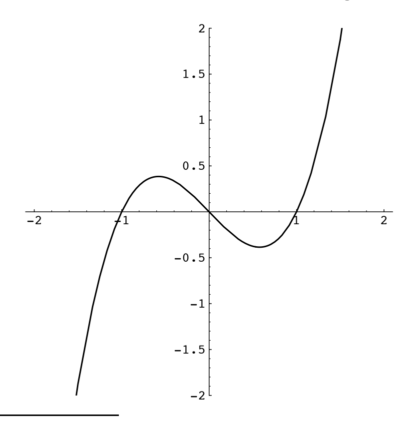
$$\Big\{\big(x, f(x)\big) \in X \times Y \mid x \in X\Big\}.$$

En los casos de funciones $f : A \subset \mathbb{R}^m \longrightarrow \mathbb{R}^n$, donde $m+n \leq 3$ la gráfica de f tiene una interpretación en el plano o espacio euclídeo intuitivos, y esta representación permite reconocer rápidamente las características de f. El resultado más importante por lo que a la continuidad se refiere es el siguiente:

Teorema 1.67 *Si $f : X \longrightarrow Y$ es una aplicación continua, entonces la aplicación $x \mapsto \big(x, f(x)\big)$ es un homeomorfismo entre X y la gráfica de f.*

Demostración: La aplicación es obviamente biyectiva y continua. Su inversa es la restricción a la gráfica de la proyección sobre el primer factor de $X \times Y$, luego también es continua. ■

Ejemplo La gráfica del polinomio $x^3 - x$ es la siguiente:



[1]Observar que desde el punto de vista de la teoría de conjuntos la gráfica de una función f es la propia función f como conjunto.

El lector debería preguntarse cómo se sabe que la gráfica de f tiene esta forma y no otra. Más adelante veremos cómo determinar analíticamente las características de la gráfica de una función, pero de momento nos bastará con lo siguiente:

Para obtener una figura como la anterior, basta programar a un ordenador para que calcule la función considerada sobre los suficientes números racionales, digamos sobre los números de la forma $k/100$, donde k varía entre -200 y 200, y dibuje un pequeño cuadrado con coordenadas $\big(x, f(x)\big)$. El resultado es una gráfica como la que hemos mostrado. El proceso sólo involucra la aritmética de los números racionales, que no tiene ninguna dificultad.

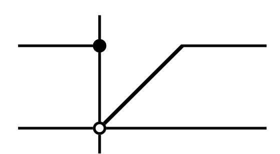
Vemos que la gráfica de f es una línea ondulada. Podemos considerarla como una imagen "típica" de espacio homeomorfo a $\mathbb{R}$. Se obtiene deformando la recta "elásticamente", sin cortarla ni pegarla. La aplicación f no es un homeomorfismo, la gráfica muestra cómo transforma a $\mathbb{R}$ en su imagen: ésta resulta de "aplastar" la curva sobre el eje vertical, con lo que $\mathbb{R}$ se "pliega" sobre sí mismo, de modo que parte de sus puntos se superponen tres a tres. ■

Veamos ahora un ejemplo de gráfica de una función discontinua.

Ejemplo Consideremos la función $f : \mathbb{R} \longrightarrow [0, 1]$ dada por

$$f(x) = \begin{cases} 1 & \text{si } x \leq 0 \\ x & \text{si } 0 < x < 1 \\ 1 & \text{si } x \geq 1 \end{cases}$$

Su gráfica es la siguiente:



Es claro que f es continua en todo punto distinto de 0. En efecto, su restricción al abierto $]-\infty, 0[$ es constante, luego continua, su restricción al abierto $]0, +\infty[$ también es continua, pues este abierto es a su vez unión de dos cerrados, $]0, 1]$ y $[1, +\infty[$, en los cuales f es continua, pues en el primero es el polinomio x y en el segundo es constante. Es fácil ver que f no es continua en 0.

La gráfica de f no es homeomorfa a $\mathbb{R}$. No estamos en condiciones de probarlo ahora. Es fácil ver que $x \mapsto \big(x, f(x)\big)$ no es un homeomorfismo, pero esto no prueba que no exista otra biyección que sí lo sea. De todos modos, intuitivamente vemos que la gráfica está formada por dos piezas, por lo que para transformar $\mathbb{R}$ en la gráfica es necesario "cortar" por algún punto. ■

Ejemplo Los homeomorfismos no pueden "cortar" ni "pegar", pero sí "estirar" arbitrariamente un conjunto. Por ejemplo, la homotecia $f(x) = ax$, donde $a > 0$ transforma el intervalo $[-1, 1]$ en el intervalo $[-a, a]$. Las traslaciones

$x \mapsto x + b$ son claramente homeomorfismos de $\mathbb{R}$, luego combinando homotecias y traslaciones podemos construir un homeomorfismo entre cualquier par de intervalos cerrados de la forma $[a, b]$. También es claro que cualquier par de intervalos abiertos $]a, b[$ son homeomorfos entre sí, al igual que cualquier par de intervalos semiabiertos $[a, b[$ y $]a, b]$. En la sección siguiente veremos que también es posible "estirar infinitamente" un intervalo, de modo que, por ejemplo, $]0, 1[$ es homeomorfo a $\mathbb{R}$. ■

La topología euclídea Notemos que toda aplicación lineal $f : \mathbb{K}^n \longrightarrow \mathbb{K}^m$ es continua, pues es cada una de sus funciones coordenadas es un polinomio. En particular todo automorfismo de $\mathbb{K}^n$ es un homeomorfismo. Si V es cualquier $\mathbb{K}$-espacio vectorial de dimensión finita n, existe un isomorfismo $f : V \longrightarrow \mathbb{K}^n$. Podemos considerar la topología en V formada por los conjuntos $f^{-1}[G]$, donde G es abierto en $\mathbb{K}^n$ para la topología euclídea. Es claro que V es así un espacio topológico homeomorfo a $\mathbb{K}^n$. Además, la topología en V no depende de la elección de f, pues si $g : V \longrightarrow \mathbb{K}^n$ es cualquier otro isomorfismo y G es un abierto en $\mathbb{K}^n$, entonces $g^{-1}[G] = f^{-1}\big[(g^{-1} \circ f)[G]\big]$ y $(g^{-1} \circ f)[G]$ es un abierto en $\mathbb{K}^n$, porque $g^{-1} \circ f$ es un isomorfismo y por consiguiente un homeomorfismo.

Más aún, la aplicación $\| \ \| : V \longrightarrow \mathbb{R}$ dada por $\|v\| = \|f(v)\|$ es claramente una norma en V que induce la topología que acabamos de definir. Esto justifica las definiciones siguientes:

Definición 1.68 Un *isomorfismo topológico* entre dos espacios vectoriales topológicos es una aplicación entre ambos que sea a la vez isomorfismo y homeomorfismo. Si V es un $\mathbb{K}$-espacio vectorial de dimensión finita n, llamaremos *topología euclídea* en V a la única topología respecto a la cual todos los isomorfismos de V en $\mathbb{K}^n$ son topológicos.

Si $h : V \longrightarrow W$ es una aplicación lineal entre dos espacios vectoriales de dimensión finita, entonces es continua, pues considerando isomorfismos (topológicos) $f : V \longrightarrow \mathbb{K}^m$ y $g : W \longrightarrow \mathbb{K}^n$ tenemos que la aplicación $h' = f^{-1} \circ h \circ g : \mathbb{K}^m \longrightarrow \mathbb{K}^n$ es lineal, luego continua, luego $h = f \circ h' \circ g^{-1}$ también es continua.

Si W es un subespacio de V, entonces la restricción a W de la topología euclídea en V es la topología euclídea. Para probarlo descomponemos $V = W \oplus W'$ y observamos que la identidad de W con la topología euclídea a W con la topología inducida es continua por ser lineal y su inversa es continua por ser la restricción de la proyección de V en W, que también es lineal. Por lo tanto se trata de un homeomorfismo y ambas topologías coinciden.

Similarmente se prueba que la topología euclídea en un producto de espacios vectoriales coincide con el producto de las topologías euclídeas.

Todo subespacio W de un espacio vectorial de dimensión finita V es cerrado. Para probarlo observamos que W puede expresarse como el núcleo de una aplicación lineal de W en $\mathbb{K}^n$, es decir, como la antiimagen de 0 por una aplicación continua y, como $\{0\}$ es cerrado, su antiimagen también.

Todos estos hechos se trasladan sin dificultad a los espacios afines sobre $\mathbb{K}$. Se define la topología euclídea en un espacio afín de modo que todas las afinidades son continuas y las variedades afines son cerradas.

***Ejemplo: La topología proyectiva** Un espacio proyectivo sobre $\mathbb{K}$ es de la forma $X = \mathrm{P}(V)$, donde V es un $\mathbb{K}$-espacio vectorial de dimensión $n+1$. Consideremos la aplicación $P : V \setminus \{\vec{0}\} \longrightarrow X$ dada por $P(\vec{v}) = \langle \vec{v} \rangle$. Vamos a considerar en X la mayor topología que hace continua a P, es decir, los abiertos de X serán los conjuntos $G \subset X$ tales que $P^{-1}[G]$ es abierto en V (respecto a la topología euclídea). Es fácil ver que los conjuntos así definidos determinan realmente una topología en X. En lo sucesivo consideraremos a todos los espacios proyectivos como espacios topológicos con esta topología. La llamaremos *topología proyectiva.* Vamos a probar algunos hechos en torno a ella:

La proyección $P : V \setminus \{\vec{0}\} \longrightarrow X$ es continua, abierta y suprayectiva.

Sólo hemos de comprobar que es abierta. Sea G un abierto en $V \setminus \{\vec{0}\}$. Hemos de ver que $G' = P^{-1}\big[P[G]\big]$ es abierto en $V \setminus \{\vec{0}\}$. Es claro que $G \subset G'$. Sea $\vec{v} \in G$. Esto significa que existe un $\vec{w} \in G$ tal que $P(\vec{v}) = P(\vec{w})$, luego $\vec{v} = \alpha\vec{w}$, para un $\alpha \in \mathbb{K}$. La homotecia $h(\vec{x}) = \alpha\vec{x}$ es un homeomorfismo que transforma G en un entorno de $\vec{v}$. Además $h[G] \subset h[G'] = G'$, luego G' es entorno de $\vec{v}$.

Las homografías entre espacios proyectivos son homeomorfismos.

Dada una homografía $H : X \longrightarrow Y$, sea $h : V \longrightarrow W$ el isomorfismo que la induce, que de hecho es un homeomorfismo. Sean $P : V \setminus \{\vec{0}\} \longrightarrow X$ y $P' : W \setminus \{\vec{0}\} \longrightarrow Y$ las proyecciones que definen la topología. Entonces, si G es abierto en Y, por definición $P'^{-1}[G]$ es abierto en $W \setminus \{\vec{0}\}$, luego $h^{-1}\big[P'^{-1}[G]\big]$ es abierto en $V \setminus \{\vec{0}\}$, pero es fácil ver que este conjunto coincide con $P^{-1}\big[H^{-1}[G]\big]$, luego $H^{-1}[G]$ es abierto, y el mismo razonamiento se aplica a la homografía inversa.

Si E es un hiperplano de V que no pasa por $\vec{0}$, entonces $P[E]$ es abierto en X y $P|_E : E \longrightarrow P[E]$ es un homeomorfismo.

Claramente $P|_E$ es inyectiva y continua. Basta probar que si A es abierto en E entonces $P[A]$ es abierto en X. Sea $B = P^{-1}\big[P[A]\big]$. Basta ver que B es abierto en $V \setminus \{\vec{0}\}$. Es claro que $B = \{\alpha\vec{v} \mid \alpha \in \mathbb{K} \setminus \{0\}, \vec{v} \in A\}$.

Escogiendo una base adecuada $\vec{v}_1, \ldots, \vec{v}_{n+1}$ en V podemos suponer que E está formado por los vectores con última coordenada es $x_{n+1} = 1$. Para cada $\vec{v} \in V$ sea $\alpha(\vec{v})$ su última coordenada. La aplicación α es continua. Sea G el conjunto de vectores de V con $\alpha(\vec{v}) \neq 0$. Tenemos que G es abierto, pues su complementario es $\alpha^{-1}\big[\{0\}\big]$. La aplicación $G \longrightarrow E$ dada por $\vec{v} \mapsto \alpha(\vec{v})^{-1}\,\vec{v}$ es continua y B es la antiimagen de A por esta aplicación, luego es abierto en G, luego en $V \setminus \{\vec{0}\}$.

Las subvariedades proyectivas de X son cerradas.

Dado un hiperplano E en V que no pase por 0, tenemos que el complementario de $P[E]$ es un hiperplano de X y, según lo que acabamos de probar, es cerrado. Como existe una homografía que transforma un hiperplano en otro cualquiera, concluimos que todos los hiperplanos son cerrados y, como toda subvariedad proyectiva es intersección de un número finito de hiperplanos, resulta que todas las subvariedades de X son cerradas.

La topología inducida por X en una subvariedad proyectiva Y es la topología proyectiva de Y.

Sea $Y = \mathrm{P}(W)$. Tomemos un punto $y \in Y$. Sea Π un hiperplano en X que no contenga a y, sea $\Pi' = Y \cap \Pi$. Entonces una base de entornos de y para las topologías de X e Y son los abiertos que contienen a y y están contenidos en $X \setminus \Pi$ e $Y \setminus \Pi'$ respectivamente. Las topologías de estos espacios son las euclídeas, luego una es la inducida por la otra, luego los entornos básicos de y en Y son las intersecciones con Y de los entornos básicos de y en X. Por consiguiente, la topología proyectiva y la topología inducida tienen una misma base de entornos de cada punto, luego son iguales.

Para terminar describiremos la topología de la recta proyectiva compleja. Tenemos que $\mathrm{P}^1(\mathbb{C}) = \mathbb{C} \cup \{\infty\}$. La topología en $\mathbb{C}$ es la euclídea. Sólo falta determinar los entornos de ∞. Como la aplicación $1/z$ es un homeomorfismo, una base de entornos de ∞ la forman las imágenes por $1/z$ de los elementos de una base de entornos de 0, por ejemplo las bolas abiertas euclídeas de centro 0. Pero la imagen de $B_\epsilon(0)$ está formada por ∞ y los puntos $z \in \mathbb{C}$ tales que $|z| > 1/\epsilon$, luego una base de entornos abiertos de ∞ la forman los complementarios de las bolas cerradas de centro 0. Es fácil ver entonces que un conjunto U es un entorno de ∞ si y sólo si $\infty \in U$ y $U \setminus \{\infty\}$ está acotado. Esta misma descripción vale para los entornos de ∞ en $\mathrm{P}^1(\mathbb{R})$.

Ejercicio: Probar que las homografías entre cónicas y rectas son homeomorfismos.

Ejemplo: La proyección estereográfica Consideremos la esfera de centro $(0, 0, 0)$ y radio 1, es decir, el conjunto $S = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$. Sea $N = (0, 0, 1)$ el "polo norte" de S.

La proyección estereográfica es la biyección entre $S \setminus \{N\}$ y $\mathbb{R}^2$ que a cada punto P de S le asigna el punto donde la recta NP corta al plano $z = 0$ (que podemos identificar con $\mathbb{R}^2$). Si P tiene coordenadas (x, y, z), la recta NP está formada por los puntos $(0, 0, 1) + \lambda(x, y, x - 1)$, con $\lambda \in \mathbb{R}$. El valor de λ que anula la tercera coordenada es el que cumple $1 + \lambda(z - 1) = 0$, o sea, $\lambda = 1/(1 - z)$, luego la proyección de P es el punto

$$f(x, y, z) = \left(\frac{x}{1-z}, \frac{y}{1-z}\right).$$

Similarmente se calcula la proyección inversa, que es

$$g(u, v) = \left(\frac{2u}{u^2+v^2+1}, \frac{2v}{u^2+v^2+1}, \frac{u^2+v^2-1}{u^2+v^2+1}\right).$$

Es evidente que tanto f como g son continuas, luego la proyección estereográfica es un homeomorfismo. Así pues, el plano es homeomorfo a una esfera menos un punto.

Equivalentemente, podemos decir que la esfera es homeomorfa al plano más un punto (extendiendo adecuadamente la topología). Conviene identificar el plano con el cuerpo complejo $\mathbb{C}$. Añadamos a $\mathbb{C}$ un punto cualquiera que representaremos por ∞, con lo que formamos el conjunto $\mathbb{C}^\infty = \mathbb{C} \cup \{\infty\}$. Podemos extender la proyección estereográfica a una biyección $f : S \longrightarrow \mathbb{C}^\infty$ sin más que asignar $f(N) = \infty$. Es obvio que si tomamos como abiertos en $\mathbb{C}^\infty$ las imágenes por f de los abiertos de S obtendremos una topología en $\mathbb{C}^\infty$ que induce en $\mathbb{C}$ la topología euclídea. Vamos a dar una descripción directa de una base de entornos de ∞.

Para ello consideramos una base de entornos de N en S, concretamente la formada por las bolas de radio menor que 1. Un entorno básico de N está formado por los puntos $(x, y, x) \in S$ tales que $\sqrt{x^2 + y^2 + (z-1)^2} < \epsilon < 1$, y como $x^2+y^2+z^2 = 1$, esto equivale a $z > 1-\epsilon^2/2$. Puesto que ϵ es arbitrario, los entornos básicos están formados por los puntos $(x, y, z) \in S$ tales que $1 - z < \epsilon$, para $\epsilon > 0$. Calculemos la imagen de uno de estos entornos.

Para ello notamos que si $(x, y, z) \neq N$,

$$|f(x, y, z)| = \sqrt{\frac{1+z}{1-z}},$$

y que si $z < z'$ entonces $z - zz' < z' - zz'$, luego

$$\frac{z}{1-z} < \frac{z'}{1-z'}, \quad \sqrt{1+\frac{2z}{1-z}} < \sqrt{1+\frac{2z'}{1-z'}}, \quad \sqrt{\frac{1+z}{1-z}} < \sqrt{\frac{1+z'}{1-z'}}.$$

En otras palabras, cuanto mayor es z, mayor es la distancia a 0 de $f(x, y, z)$. Por consiguiente, los puntos tales que $z > 1 - \epsilon$ se corresponden con los puntos tales que

$$|f(x, y, z)| > |f(x, y, 1-\epsilon)| = \sqrt{\frac{2}{\epsilon} - 1}.$$

Ahora bien, cualquier $M > 0$ es de esta forma para algún épsilon, concretamente $\epsilon = 2/(M^2 + 1)$. Concluimos que los entornos básicos de ∞ son los conjuntos de la forma

$$\{z \in \mathbb{C} \mid |z| > M\} \cup \{\infty\}.$$

Esto significa que un punto del plano está más cerca de infinito cuanto más lejos está de 0 (de hecho, cuanto más lejos está de cualquier punto concreto del plano). ■

***Nota** Obviamente $\mathbb{C}^\infty$ es la recta proyectiva compleja con la topología proyectiva. Otra forma de llegar al mismo resultado es la siguiente: Si consideramos la proyección estereográfica entre S y la recta proyectiva compleja, podemos expresarla como $f(x, y, z) = \langle(x, y, 1-z)\rangle$, es decir, como la composición de la aplicación continua $(x, y, z) \mapsto (x, y, 1-z)$ con la aplicación continua $(x, y, z) \mapsto \langle(x, y, z)\rangle$, luego es continua. Podríamos probar también que

la inversa es continua, pero no merece la pena el esfuerzo, pues en el capítulo siguiente (ver 2.12) será evidente a partir de lo que ya hemos probado. La conclusión es, pues, que la proyección estereográfica es un homeomorfismo entre la esfera con la topología euclídea y la recta compleja con la topología proyectiva.

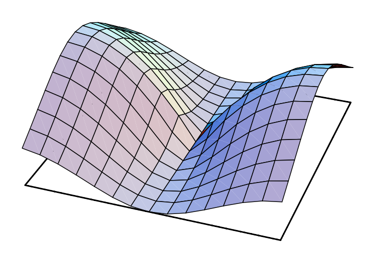
De aquí se sigue fácilmente que si dotamos al plano hiperbólico de la topología relativa (es decir, si dotamos al plano de Klein con la topología euclídea) entonces la topología correspondiente en el círculo y en el semiplano de Poincaré es también la euclídea. Teniendo en cuenta que los círculos de centro 0 en el plano de Klein coinciden con los euclídeos (y que las isometrías son obviamente homeomorfismos) es claro que la esta topología es la inducida por la métrica hiperbólica. ■

1.6 Límites de funciones

El concepto de límite es, junto al de continuidad, uno de los conceptos más importantes a los que la estructura topológica sirve de soporte. Para comprender su contenido podemos considerar la función $f : \mathbb{R}^2 \setminus \{(0,0)\} \longrightarrow \mathbb{R}$ dada por

$$f(x,y) = x^2\left(\frac{2}{\sqrt{x^2+y^2}} - 1\right).$$

Esta expresión no tiene sentido cuando $(x,y) = (0,0)$, por lo que es natural preguntarse si la gráfica de f mostrará alguna particularidad que explique por qué no puede ser calculada en este punto. He aquí dicha gráfica:



Su aspecto es el de una superficie homeomorfa a $\mathbb{R}^2$, pero sabemos que no está definida en $(0,0)$. De hecho la gráfica hace pensar que $f(0,0) = 0$. La explicación es que la función $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$ definida mediante

$$f(x,y) = \begin{cases} x^2\left(\frac{2}{\sqrt{x^2+y^2}} - 1\right) & \text{si } (x,y) \neq (0,0), \\ 0 & \text{si } (x,y) = (0,0) \end{cases}$$

es continua, aunque esto no es evidente y, de hecho, no estamos en condiciones de probarlo ahora. Si queremos expresar la situación en términos de la expresión original de f, no definida en $(0,0)$, habremos de decir que f transforma los

puntos de alrededor de $(0,0)$ en puntos de alrededor de 0, o también, que los valores que toma f son más cercanos a 0 cuanto más cercanos a $(0,0)$ son los puntos que consideramos. Todo esto tiene sentido aunque la función f no esté definida en $(0,0)$.

Comenzamos introduciendo el concepto topológico que permite expresar con rigor las ideas que acabamos de introducir:

Definición 1.69 Sean X, Y espacios topológicos, $A \subset X$ y $f : A \longrightarrow Y$. Sea $a \in A'$ y $b \in Y$. Diremos que f *converge* a b cuando x tiende a a si para todo entorno V de b existe un entorno U de a tal que si $x \in U \cap A$ y $x \neq a$, entonces $f(x) \in V$.

La interpretación es clara: los puntos $f(x)$ están alrededor de b [= en un entorno arbitrario V de b] siempre que x está alrededor de a [= en un entorno adecuado U de a, que dependerá de V], o más simplemente: si f envía los puntos de alrededor de a a los alrededores de b.

Si Y es un espacio de Hausdorff, una función converge a lo sumo a un único b para cada punto a. En efecto, si f converge a dos puntos b y b' cuando x tiende a a, podríamos tomar entornos disjuntos V y V' de b y b', para los cuales existirían entornos U y U' de a de modo que si $x \in U \cap U' \cap A$, entonces $f(x) \in V \cap V'$, contradicción.

Por ello, si se da la convergencia, diremos que b es el *límite* cuando x tiende a a de $f(x)$, y lo representaremos por

$$b = \lim_{x \to a} f(x).$$

No exigimos que la función f esté definida en a. Tan sólo que a sea un punto de acumulación del dominio de f o, de lo contrario, no existirían puntos x a los que aplicar la definición y f convergería trivialmente a todos los puntos de Y.

En estos términos, lo que afirmábamos antes es que existe

$$\lim_{(x,y)\to(0,0)} x^2 \left(\frac{2}{\sqrt{x^2+y^2}} - 1 \right) = 0,$$

de modo que los valores que toma esta expresión se acercan más a 0 cuanto más se acercan las variables al punto $(0,0)$. Todavía no podemos probarlo.

Por supuesto es posible que la función f esté definida en a, pero esto es irrelevante, pues en la definición de límite aparecen sólo puntos $x \neq a$, lo que significa que el límite es independiente de $f(a)$. En otras palabras, si modificáramos el valor de $f(a)$, la existencia del límite y su valor concreto no se alterarían.

También es obvio que la existencia o no de límite sólo depende del comportamiento de la función en un entorno del punto. En otras palabras, que si dos funciones coinciden en un entorno de un punto a (salvo quizá en a) entonces ambas tienen límite en a o ninguna lo tiene y, si lo tienen, éstos coinciden.

La relación entre los límites y la continuidad es la siguiente:

Teorema 1.70 *Sean X, Y espacios topológicos y $f : X \longrightarrow Y$. Sea $a \in X'$. Entonces f es continua en a si y sólo si existe $\lim\limits_{x\to a} f(x) = f(a)$.*

Demostración: Si el límite vale $f(a)$, entonces la definición de límite se cumple trivialmente cuando $x = a$ y lo que queda es la definición de continuidad en a. Recíprocamente, la definición de continuidad de f en a implica trivialmente la definición de límite con $b = f(a)$. ■

Mediante este teorema podemos deducir las propiedades algebraicas de los límites a partir de las propiedades correspondientes de las funciones continuas.

Teorema 1.71 *Sean f, $g : A \subset X \longrightarrow \mathbb{K}$ dos funciones definidas sobre un espacio topológico X y sea $a \in A'$. Si existen*

$$\lim_{x\to a} f(x) \quad y \quad \lim_{x\to a} g(x)$$

entonces también existen

$$\begin{aligned}
\lim_{x\to a}\big(f(x)+g(x)\big) &= \lim_{x\to a} f(x) + \lim_{x\to a} g(x),\\
\lim_{x\to a} f(x)g(x) &= \Big(\lim_{x\to a} f(x)\Big)\Big(\lim_{x\to a} g(x)\Big),\\
\lim_{x\to a}\frac{f(x)}{g(x)} &= \frac{\lim\limits_{x\to a} f(x)}{\lim\limits_{x\to a} g(x)},
\end{aligned}$$

(suponiendo, además, en el tercer caso que $\lim\limits_{x\to a} g(x) \neq 0$.)

Demostración: La prueba es la misma en todos los casos. Veamos la primera. Consideramos la función f' que coincide con f en todos los puntos salvo en a, donde toma el valor del límite. Definimos igualmente g'. Entonces el teorema anterior implica que f' y g' son continuas en a, luego $f'+g'$ también lo es, luego por el teorema anterior existe

$$\lim_{x\to a}\big(f(x)+g(x)\big) = f'(a) + g'(a),$$

pero como $f'+g'$ coincide con $f+g$ salvo en a, el límite de $f'+g'$ coincide con el de $f+g$, y tenemos la relación buscada. ■

Es claro que el resultado sobre suma de límites es válido cuando las funciones toman valores en cualquier espacio vectorial topológico. Además en tal caso al multiplicar una función por un escalar el límite se multiplica por el mismo (para el caso de $\mathbb{K}$ esto es un caso particular de la segunda igualdad). La misma técnica nos da inmediatamente:

Teorema 1.72 *Sea $f : A \subset X \longrightarrow X_1 \times \cdots \times X_n$ una aplicación entre espacios topológicos, que será de la forma $f(x) = \big(f_1(x_1), \ldots, f_n(x)\big)$, para ciertas funciones $f_i : X \longrightarrow X_i$. Sea $a \in A'$. Entonces existe $\lim\limits_{x\to a} f(x)$ si y sólo si existen todos los límites $\lim\limits_{x\to a} f_i(x)$ y en tal caso*

$$\lim_{x\to a} f(x) = \Big(\lim_{x\to a} f_1(x), \ldots, \lim_{x\to a} f_n(x)\Big).$$

Para calcular el límite que tenemos pendiente necesitamos un hecho que a menudo resulta útil. Diremos que una función f con valores en un espacio métrico está *acotada* si su imagen es un conjunto acotado.

Teorema 1.73 *Sean f, $g : A \subset X \longrightarrow E$ dos funciones definidas de un espacio topológico X en un espacio normado E y sea $a \in A'$. Si existe $\lim\limits_{x\to a} f(x) = 0$ y g está acotada, entonces existe $\lim\limits_{x\to a} f(x)g(x) = 0$.*

Demostración: Si g está acotada, existe un $M > 0$ tal que $\|g(x)\| \leq M$ para todo $x \in A$. Entonces $\|f(x)g(x)\| \leq M\|f(x)\|$. El hecho de que f tienda a 0, sustituyendo los entornos en E de la definición general por bolas abiertas, queda así:

Para todo $\epsilon > 0$ existe un entorno U de a tal que si $x \in U \cap A$ y $x \neq a$, entonces $\|f(x)\| < \epsilon$.

Tomamos ahora $\epsilon > 0$ y aplicamos este hecho a ϵ/M, con lo que si $x \in U \cap A$ y $x \neq a$, tenemos que $\|f(x)g(x)\| \leq M\|f(x)\| < \epsilon$, y esto significa que fg tiende a 0. ■

Ejemplo Se cumple

$$\lim_{(x,y)\to(0,0)} x^2 \left(\frac{2}{\sqrt{x^2+y^2}} - 1 \right) = 0,$$

En efecto, basta probar que

$$\lim_{(x,y)\to(0,0)} \frac{2x^2}{\sqrt{x^2+y^2}} = 0,$$

pues el otro sumando, x^2 tiende obviamente a 0, por continuidad. Factorizamos

$$x \frac{2x}{\sqrt{x^2+y^2}}.$$

El primer factor tiende a 0 y el segundo está acotado, pues se comprueba fácilmente que

$$-1 \leq \frac{x}{\sqrt{x^2+y^2}} \leq 1.$$

Ahora basta aplicar el teorema anterior. ■

El hecho de que la composición de funciones continuas es continua se traduce ahora en el hecho siguiente:

Teorema 1.74 *Sea $f : X \longrightarrow Y$ y $g : Y \longrightarrow Z$, sea $a \in X'$ y supongamos que existe*

$$\lim_{x\to a} f(x) = b$$

y que g es continua en b. Entonces

$$g\Big(\lim_{x\to a} f(x)\Big) = \lim_{x\to a} g\big(f(x)\big).$$

DEMOSTRACIÓN: Sea U un entorno de $g(b)$. Entonces $g^{-1}[U]$ es un entorno de b. Existe un entorno V de a tal que si $x \in V$, $x \neq a$, entonces $f(x)$ in $g^{-1}[U]$, luego $g(f(x)) \in U$. Por lo tanto se cumple la definición de límite. ■

Por ejemplo, la continuidad de la raíz cuadrada implica que

$$\lim_{(x,y)\to(0,0)} |x| \sqrt{\frac{2}{\sqrt{x^2+y^2}} - 1} = 0.$$

Ejercicio: Dar un ejemplo que muestre la falsedad de la afirmación siguiente: Dadas dos funciones $f,\ g : \mathbb{R} \longrightarrow \mathbb{R}$, si existen $\lim\limits_{x\to a} f(x) = b$, $\lim\limits_{x\to b} g(x) = c$ entonces existe también $\lim\limits_{x\to a} g(f(x)) = c$. Probar que es cierta si $f(x) \neq b$ para $x \neq a$.

Limites restringidos El límite de una función en un punto depende del dominio de ésta, por eso es importante lo que ocurre al restringir una función a dominios menores. Por ejemplo pensemos en la función $f : \mathbb{R} \longrightarrow \mathbb{R}$ dada por

$$f(x) = \begin{cases} -1 & \text{si } x < 0 \\ 1 & \text{si } x \geq 0 \end{cases}$$

Es claro que no tiene límite en 0, pero sí lo tienen las funciones $f|_{]-\infty,0[}$ y $f|_{]0,+\infty[}$. Si consideramos sólo la parte de la izquierda de la función, resulta que es constante igual a -1, de donde su límite es -1. Igualmente el límite de la parte derecha es 1. Por ello definimos:

Definición 1.75 Sea $f : A \subset X \longrightarrow Y$, $B \subset A$ y $a \in B'$. Definimos

$$\lim_{\substack{x\to a\\ B}} f(x) = \lim_{x\to a} f|_B(x).$$

Para el caso de funciones definidas sobre subconjuntos de $\mathbb{R}$ se define

$$\lim_{x\to a^-} f(x) = \lim_{\substack{x\to a\\ ]-\infty,a[}} f(x), \qquad \lim_{x\to a^+} f(x) = \lim_{\substack{x\to a\\ ]a,+\infty[}} f(x),$$

y se llaman, respectivamente, límite por la izquierda y límite por la derecha de f en a. También se les llama *límites laterales.* Su utilidad se debe al teorema siguiente:

Teorema 1.76 *Sea A un subconjunto de un espacio X y $f : A \longrightarrow Y$. Supongamos que $A = B_1 \cup \cdots \cup B_n$ y que a es un punto de acumulación de todos estos conjuntos. Entonces existe $\lim\limits_{x\to a} f(x)$ si y sólo si existen los límites $\lim\limits_{\substack{x\to a\\ B_i}} f(x)$ para $i = 1, \ldots, n$ y todos coinciden. En tal caso $f(x)$ es igual al límite común.*

DEMOSTRACIÓN: Si existen tales límites y todos valen L, sea V un entorno de L. Por definición existen entornos U_i de a tales que si $x \in U_i \cap (A \cap B_i)$ y $x \neq a$, entonces $f(x) \in V$. Si U es la intersección de los conjuntos U_i, entonces U es un entorno de a y si $x \in U \cap A$, $x \neq a$, existirá un i tal que $x \in B_i$, luego $f(x) \in V$, es decir, $\lim\limits_{x\to a} f(x) = L$. El recíproco es más sencillo. ■

Ejemplo Se cumple

$$\lim_{(x,y)\to(0,0)} x\sqrt{\frac{2}{\sqrt{x^2+y^2}}-1}=0.$$

Para comprobarlo calculamos los límites

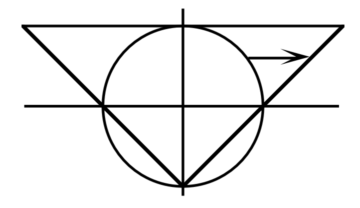
$$\lim_{\substack{(x,y)\to(0,0)\\ x\leq 0}} x\sqrt{\frac{2}{\sqrt{x^2+y^2}}-1}=-\lim_{\substack{(x,y)\to(0,0)\\ x\leq 0}} |x|\sqrt{\frac{2}{\sqrt{x^2+y^2}}-1}=0,$$

$$\lim_{\substack{(x,y)\to(0,0)\\ x\geq 0}} x\sqrt{\frac{2}{\sqrt{x^2+y^2}}-1}=\lim_{\substack{(x,y)\to(0,0)\\ x\geq 0}} |x|\sqrt{\frac{2}{\sqrt{x^2+y^2}}-1}=0,$$

y aplicamos el teorema anterior. ■

Ejemplo: la proyección cónica Veamos ahora un caso en el que nos aparece de forma natural el cálculo de un límite.

Consideremos de nuevo la esfera S y el cono C formado al unir el polo sur $(0,0-1)$ con los puntos del ecuador de S. La figura muestra un corte transversal de S y C. Vamos a probar que $S\setminus\{N\}$ es homeomorfo a una porción de C mediante la aplicación que traslada radialmente cada punto.



Un punto (x,y,z) de C está en el segmento que une el punto $(0,0,-1)$ con un punto $(a,b,0)$, donde $a^2+b^2=1$. Por tanto será de la forma

$$(x,y,z)=(0,0,-1)+\lambda(a,b,-1),\quad \text{con } \lambda\in\mathbb{R}.$$

Entonces $\lambda=z+1$, luego $x=a(z+1)$, $y=b(z+1)$. De aquí se sigue que

$$z+1=\pm\sqrt{x^2+y^2}.$$

Recíprocamente, si un punto cumple esta ecuación, si $z=-1$ ha de ser $(0,0,-1)$, que está en C, y si $z\neq -1$, entonces los valores

$$\lambda=z+1,\quad a=\frac{x}{z+1},\quad b=\frac{y}{z+1},$$

permiten expresar a (x,y,z) en la forma paramétrica anterior, luego se trata de un punto de C. Nos interesa sólo la porción de C formada por los puntos con altura entre -1 y 1, por lo que definimos

$$C=\big\{(x,y,z)\in\mathbb{R}^3 \mid z+1=\sqrt{x^2+y^2},\ -1\leq z<1\big\}.$$

Notar que los puntos con $z = 1$ no están en C. El homeomorfismo que buscamos ha de transformar cada punto (x, y, z) de $S \setminus \{N\}$ en un punto $(\lambda x, \lambda y, z)$, donde $\lambda \geq 0$ es el adecuado para llegar a C. La condición es

$$\sqrt{(\lambda x)^2 + (\lambda y)^2} = z + 1$$

y, teniendo en cuenta que los puntos de la esfera cumplen $\sqrt{x^2 + y^2} = \sqrt{1 - z^2}$,

$$\lambda = \frac{1+z}{\sqrt{x^2+y^2}} = \sqrt{\frac{1+z}{1-z}}.$$

Por lo tanto $f : S \setminus \{N\} \longrightarrow C$ será la aplicación dada por

$$f(x, y, z) = \left(\sqrt{\frac{1+z}{1-z}}\, x, \sqrt{\frac{1+z}{1-z}}\, y, z \right).$$

La inversa se calcula sin dificultad a partir de esta expresión:

$$g(x, y, z) = \left(\sqrt{\frac{1-z}{1+z}}\, x, \sqrt{\frac{1-z}{1+z}}\, y, z \right), \quad \text{si } (x, y, z) \neq (0, 0, -1),$$

y por supuesto $g(0, 0, -1) = (0, 0, -1)$.

Obviamente f es continua. Lo mismo vale para g en todos los puntos salvo en $(0, 0, -1)$. Para probar la continuidad en este punto hacemos uso de la igualdad $1 + z = \sqrt{x^2 + y^2}$, que cumplen todos los puntos de C, la cual nos permite expresar g como

$$g(x, y, z) = \left(x \sqrt{\frac{2}{\sqrt{x^2+y^2}} - 1},\ y \sqrt{\frac{2}{\sqrt{x^2+y^2}} - 1}, z \right).$$

Basta probar que

$$\lim_{(x,y,z) \to (0,0,-1)} \left(x \sqrt{\frac{2}{\sqrt{x^2+y^2}} - 1},\ y \sqrt{\frac{2}{\sqrt{x^2+y^2}} - 1}, z \right) = (0, 0, -1),$$

pero los dos primeros límites son el que hemos calculado como ejemplo a lo largo de la sección.

Así pues, f es un homeomorfismo entre $S \setminus \{N\}$ y C. Ahora bien, es inmediato comprobar que la proyección sobre el plano XY es un homeomorfismo entre C y la bola abierta de centro 0 y radio 2 (la aplicación inversa es $(x, y) \mapsto (x, y, -1 + \sqrt{x^2 + y^2})$, claramente continua), con lo que la composición

$$h(x, y, z) = \left(\sqrt{\frac{1+z}{1-z}}\, x, \sqrt{\frac{1+z}{1-z}}\, y \right).$$

es un homeomorfismo entre $S \setminus \{N\}$ y dicha bola. Más aún, si componemos la inversa de la proyección estereográfica con esta aplicación obtenemos un homeomorfismo entre $\mathbb{R}^2$ y el disco abierto. Es fácil ver que la composición es

$$t(x,y) = \left(\frac{2x\sqrt{x^2+y^2}}{x^2+y^2+1}, \frac{2y\sqrt{x^2+y^2}}{x^2+y^2+1} \right).$$

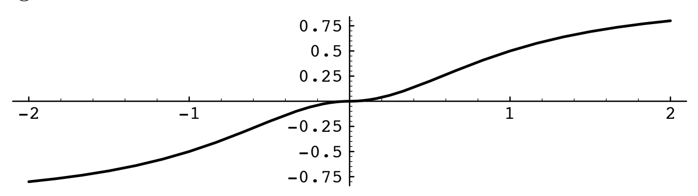
Si quitamos los doses obtenemos un homeomorfismo $t : \mathbb{R}^2 \longrightarrow D$, donde D es la bola abierta (euclídea) de centro 0 y radio 1. Concretamente:

$$t(x,y) = \left(\frac{x\sqrt{x^2+y^2}}{x^2+y^2+1}, \frac{y\sqrt{x^2+y^2}}{x^2+y^2+1} \right).$$

Si restringimos esta aplicación a $\mathbb{R}$, es decir, a los puntos $(x, 0)$ obtenemos un homeomorfismo entre $\mathbb{R}$ y el intervalo $]-1, 1[$. Concretamente

$$t(x) = \frac{x|x|}{x^2+1}.$$

He aquí su gráfica:



Si lo restringimos a $]0, +\infty[$ obtenemos un homeomorfismo entre este intervalo y $]0, 1[$. A saber:

$$t(x) = \frac{x^2}{x^2+1}.$$

A partir de aquí es fácil ver que dos intervalos abiertos cualesquiera (acotados o no) son homeomorfos. ■

Límites infinitos Consideremos ahora límites de funciones con valores en $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$ o en $\mathbb{C}^\infty = \mathbb{C} \cup \{\infty\}$. Recordando que los entornos básicos de $+\infty$ son los intervalos $]M, +\infty]$, es claro que

$$\lim_{x \to a} f(x) = +\infty$$

significa que para todo $M > 0$ existe un entorno U de a tal que si $x \in U$, $x \neq a$ está en el dominio de f, entonces $f(x) > M$. Similarmente ocurre con $-\infty$. Que el límite valga ∞ significa que para todo $M > 0$ existe un entorno U de a tal que si $x \in U$, $x \neq a$ está en el dominio de f, entonces $|f(x)| > M$.

Es fácil probar que los teoremas del estilo de "el límite de la suma es la suma de los límites" etc. valen también en los casos siguientes:

$$+\infty + (+\infty) = +\infty, \quad -\infty + (-\infty) = -\infty,$$

$$+\infty + a = +\infty, \quad -\infty + a = -\infty, \quad \infty + a = \infty,$$

$$(+\infty)(+\infty) = +\infty, \quad (-\infty)(-\infty) = +\infty, \quad \infty\,\infty = \infty.$$

$$\text{si } a > 0, \quad (+\infty)a = +\infty, \quad (-\infty)a = -\infty;$$

$$\text{si } a < 0, \quad (+\infty)a = -\infty, \quad (-\infty)a = +\infty.$$

$$\text{Si } a \neq 0, \quad \infty\, a = \infty, \quad \frac{a}{\pm\infty} = \frac{a}{\infty} = 0, \quad \frac{a}{0} = \infty.$$

Por ejemplo, la igualdad $+\infty + (+\infty) = +\infty$ es una forma de expresar que la suma de dos funciones $f,\, g : A \subset X \longrightarrow \overline{\mathbb{R}}$ que tienden a $+\infty$ en un punto $a \in A'$, tiende también a $+\infty$. La prueba es sencilla: dado un $M > 0$ existen entornos U y V de a tales que si $x \in U \cap A'$ entonces $f(x) > M/2$ y si $x \in V \cap A'$ entonces $g(x) > M/2$, con lo que si $x \in U \cap V \cap A$ entonces $f(x) + g(x) > M$, luego se cumple la definición de límite. Similarmente se prueban todas las demás.

Cálculo de límites Estudiamos ahora los límites más simples y frecuentes. Comencemos con los límites en infinito de los polinomios. Obviamente

$$\lim_{x \to \pm\infty} x = \pm\infty.$$

Aplicando n veces la regla $(+\infty)(+\infty) = +\infty$ vemos que si $n > 0$ entonces

$$\lim_{x \to +\infty} x^n = +\infty.$$

El límite en $-\infty$ será obviamente $(-1)^n\,\infty$. Si $n < 0$ usamos la regla $1/\pm\infty = 0$ y si $n = 0$ tenemos que x^0 es la constante 1, luego en total:

$$\lim_{x \to +\infty} x^n = \begin{cases} +\infty & \text{si } n > 0, \\ 1 & \text{si } n = 0, \\ 0 & \text{si } n < 0. \end{cases}$$

El límite en $-\infty$ difiere sólo en el primer caso, de forma obvia.

Para calcular el límite de un polinomio multiplicamos y dividimos por la potencia de x de mayor grado:

$$\lim_{x \to +\infty} 8x^5 - 3x^4 + x^3 + 2x - 5 = \lim_{x \to +\infty} x^5 \left(8 - \frac{3}{x} + \frac{1}{x^2} + \frac{2}{x^4} - \frac{5}{x^5}\right) = +\infty.$$

En general, si $p(x) \in \mathbb{R}[x]$ no es constante, se cumple $\lim\limits_{x \to \pm\infty} P(x) = \pm\infty$, donde el signo depende en forma obvia del signo del coeficiente director de P y de la paridad del grado si tendemos a $-\infty$. En $\mathbb{C}^\infty$ todos los polinomios no constantes tienen límite ∞.

En particular vemos que, a diferencia de casos como $1/\infty$ o $+\infty + \infty$, no hay una regla general para calcular un límite de la forma $+\infty - \infty$. En efecto,

los tres límites siguientes son de este tipo y cada uno da un resultado distinto. Por ello se dice que $+\infty - \infty$ es una *indeterminación.*

$$\lim_{x\to+\infty} x^5 - x^2 = +\infty, \qquad \lim_{x\to+\infty} x^2 - x^5 = -\infty, \qquad \lim_{x\to+\infty} (x+2) - (x+1) = 1.$$

Nos ocupamos ahora de los límites de fracciones algebraicas (cocientes de polinomios). El ejemplo siguiente ilustra el caso general:

$$\lim_{x\to+\infty} \frac{6x^4 - 3x^3 + x + 1}{2x^4 + x^2 - 2x + 2} = \lim_{x\to+\infty} \frac{6 - \frac{3}{x} + \frac{1}{x^3} + \frac{1}{x^4}}{2 + \frac{1}{x^2} - \frac{2}{x^3} + \frac{2}{x^4}} = \frac{6}{2} = 3.$$

Es claro que el límite en $\pm\infty$ del cociente de dos polinomios del mismo grado es el cociente de los términos directores. Si los grados son distintos, al dividir entre la potencia de mayor grado obtenemos 0 si el grado del denominador es mayor y $\pm\infty$ si es mayor el del numerador, donde el signo se calcula de forma obvia. Esto vale igualmente (salvo lo dicho de los signos) para límites en $\mathbb{C}^\infty$.

Vemos, pues que el caso ∞/∞ es también una indeterminación.

Por ejemplo, antes hemos probado que la función

$$t(x) = \frac{x|x|}{x^2+1}$$

es un homeomorfismo entre $\mathbb{R}$ y el intervalo $]-1,1[$. Ahora es claro que

$$\lim_{x\to\pm\infty} = \pm 1,$$

luego si definimos $t(\pm\infty) = \pm 1$ tenemos una biyección continua $t : \overline{\mathbb{R}} \longrightarrow [-1,1]$. En el capítulo siguiente veremos que es un homeomorfismo.

Ejercicio: Calcular t^{-1} y probar que es continua.

1.7 Convergencia de sucesiones

Si los límites de funciones tienen especial interés en el cálculo diferencial, en topología son más importantes los límites de sucesiones, que en realidad son un caso particular de los primeros.

Definición 1.77 Una *sucesión* en un conjunto X es una aplicación $a : \mathbb{N} \longrightarrow X$. Se suele representar $\{a_n\}_{n=0}^\infty$ o, más gráficamente:

$$a_0, \quad a_1, \quad a_2, \quad a_3, \quad \ldots \quad a_n, \quad \ldots$$

En realidad, lo único que nos interesa de una sucesión es el orden en el que se suceden sus elementos, y no la numeración que éstos reciben. Por ello en la práctica admitiremos sucesiones que comiencen en índices mayores de 0. Por ejemplo, la sucesión $\{n\}_{n=0}^\infty$ y la sucesión $\{n-7\}_{n=0}^\infty$ son conjuntistamente

distintas, pero para nosotros serán la misma sucesión, la sucesión de los números naturales.

Por simplicidad, en teoría hablaremos siempre de sucesiones $\{a_n\}_{n=0}^{\infty}$, aunque en la práctica comenzaremos a partir del índice que más convenga.

La razón por la que estas diferencias no nos van a afectar es que sólo nos van a interesar las propiedades finales de las sucesiones. Diremos que una sucesión $\{a_n\}_{n=0}^{\infty}$ cumple *finalmente* una propiedad si existe un $n_0 \in \mathbb{N}$ tal que a_n cumple la propiedad para todo $n \geq n_0$.

Por ejemplo, una sucesión es *constante* si todos sus términos son iguales a un cierto elemento x. Una sucesión es *finalmente constante* si es constante a partir de un término. La sucesión siguiente es finalmente igual a 5:

$$3, \quad 1, \quad -2, \quad 8, \quad 5, \quad 5, \quad 5, \quad 5, \quad 5, \quad 5, \quad \ldots$$

Una *subsucesión* de $\{a_n\}_{n=0}^{\infty}$ es una sucesión de la forma $\{a_{n_k}\}_{k=0}^{\infty}$, donde $\{n_k\}_{k=0}^{\infty}$ es una sucesión estrictamente creciente de números naturales, es decir, si $k < k'$, entonces $n_k < n_{k'}$.

Una observación útil es que en estas condiciones siempre se cumple $k \leq n_k$. Se prueba fácilmente por inducción sobre k.

Definición 1.78 Sea X un espacio topológico, $l \in X$ y $\{a_n\}_{n=0}^{\infty}$ una sucesión en X. Diremos que a_n *converge* a l si está finalmente en todo entorno de l, o sea, si para cada entorno V de l existe un $n_0 \in \mathbb{N}$ tal que $a_n \in V$ para $n \geq n_0$.

Así pues, una sucesión converge a un punto l si está finalmente alrededor de l, es decir, si todos sus términos salvo un número finito están en cualquier entorno de l prefijado.

Si X es un espacio de Hausdorff y $\{a_n\}_{n=0}^{\infty}$ converge en X, entonces el punto al cual converge es único, pues puntos distintos tienen entornos disjuntos, y una sucesión no puede estar finalmente en dos conjuntos disjuntos. Representaremos por

$$\lim_n a_n$$

al único límite de la sucesión $\{a_n\}_{n=0}^{\infty}$ en un espacio de Hausdorff, cuando éste exista.

Veamos ahora que esta noción de límite es un caso particular del que hemos estudiado en la sección anterior. Consideremos $\mathbb{N}^{\infty} = \mathbb{N} \cup \{\infty\}$ con la topología inducida desde $\mathbb{C}^{\infty}$. Es fácil ver que la topología en $\mathbb{N}$ es la discreta y una base de entornos de ∞ la forman los conjuntos $\{n \in \mathbb{N} \mid n \geq n_0\} \cup \{\infty\}$. En estos términos, una sucesión $\{a_n\}_{n=0}^{\infty}$ converge a un punto l si y sólo si para todo entorno U de l existe un entorno V de ∞ tal que si $n \in V$, $n \neq \infty$ entonces $a_n \in U$, es decir, si y sólo si la sucesión, vista como función $\mathbb{N} \longrightarrow X$ converge a l cuando n tiende a ∞. En tal caso

$$\lim_n a_n = \lim_{n \to \infty} a_n.$$

Sabiendo esto, todos los resultados generales para límites de funciones valen para sucesiones, por ejemplo, el límite de la suma de dos sucesiones de números

reales es la suma de los límites, etc. Veamos ahora algunos hechos específicos sobre sucesiones:

Una sucesión converge si y sólo si converge finalmente, es decir, $\{a_n\}_{n=0}^{\infty}$ converge a l si y sólo si la sucesión $\{a_n\}_{n=k}^{\infty}$ converge a l. En particular, las sucesiones finalmente constantes convergen.

(Esto es consecuencia de que el límite de una función en un punto — el punto ∞ en este caso— depende sólo del comportamiento de la función en un entorno del punto).

Si $A \subset X$, $\{a_n\}_{n=0}^{\infty} \subset A$ y $l \in A$, entonces $\{a_n\}_{n=0}^{\infty}$ converge a l en A si y sólo si converge a l en X.

(Pues los entornos de l en A son las intersecciones con A de los entornos de l en X y, como la sucesión está en A, es equivalente que esté finalmente en un entorno de l en A o que esté finalmente en un entorno de l en X.)

Este hecho nos dice que la convergencia depende exclusivamente de la sucesión y de su límite, y no del espacio en el que los consideremos (siempre que no cambiemos de topología, claro está). Sin embargo, también de aquí se desprende que una sucesión convergente deja de serlo si eliminamos su límite. Por ejemplo, la sucesión $\{1/n\}_{n=0}^{\infty}$ no converge en el espacio $]0,1]$, pues si convergiera a un punto l, tendríamos que en $\mathbb{R}$ debería converger a la vez a l y a 0.

Si una sucesión converge a un punto l, entonces todas sus subsucesiones convergen a l.

Pues si $\{a_n\}_{n=0}^{\infty}$ converge a l y $\{a_{n_k}\}_{k=0}^{\infty}$ es una subsucesión, para todo entorno U de l existe un n_0 tal que si $n \geq n_0$ se cumple $a_n \in U$, pero si $k \geq n_0$ entonces $n_k \geq k \geq n_0$, luego $a_{n_k} \in U$.

Ejercicio: Demostrar que una sucesión $\{x^n\}_{n=0}^{\infty}$ converge en un espacio producto $\prod\limits_{i \in I} X_i$ a un punto x si y sólo si las sucesiones $\{x_i^n\}_{n=0}^{\infty}$ convergen a x_i para todo $i \in I$.

No podemos continuar nuestro estudio de las sucesiones sin introducir un nuevo concepto. Sucede que las sucesiones no se comportan adecuadamente en todos los espacios topológicos, sino sólo en aquellos que cumplen la siguiente propiedad adicional, por lo demás muy frecuente:

Definición 1.79 Un espacio X cumple el *primer axioma de numerabilidad* (1AN) si para todo punto $x \in X$ existe una base de entornos de x de la forma $\{V_n\}_{n=0}^{\infty}$.

Esta propiedad la tienen todos los espacios métricos, pues si x es un punto de un espacio métrico, una base de entornos de x la forman los conjuntos $\{B_{1/n}(x)\}_{n=1}^{\infty}$. Así pues, todos los espacios que manejamos cumplen 1AN. En el caso de $\mathbb{C}^{\infty}$, en lugar de considerar la métrica inducida desde la esfera, es

más fácil considerar directamente la base de entornos de ∞ formada por los conjuntos siguientes:

$$E_n = \{z \in \mathbb{C} \bigm| |z| > n\} \cup \{\infty\}, \quad \text{para } n = 1, 2, \ldots$$

En $\overline{\mathbb{R}}$, una base de entornos de $+\infty$ es $\big\{]n, +\infty]\big\}\}_{n=0}^{\infty}$ y una base de entornos de $-\infty$ es $\big\{[-\infty, n[\big\}\}_{n=0}^{\infty}$, luego este espacio también cumple 1AN.

Observemos que si X es un espacio que cumple 1AN y $x \in X$, podemos tomar una base de entornos de x de la forma $\{V_n\}_{n=0}^{\infty}$ que cumpla además

$$V_0 \quad \supset \quad V_1 \quad \supset \quad V_2 \quad \supset \quad V_3 \quad \supset \quad \ldots$$

Si una base dada $\{W_n\}_{n=0}^{\infty}$ no lo cumple, tomamos $V_n = W_0 \cap \cdots \cap W_n$ y así tenemos las inclusiones indicadas. Todas las bases de entornos de los ejemplos que acabamos de dar son decrecientes en este sentido.

Consideremos ahora esta sucesión:

$$1, \quad -1, \quad 1, \quad -1, \quad 1, \quad -1, \quad 1, \quad -1, \quad 1, \quad -1, \quad 1, \quad -1, \quad \ldots$$

Es obvio que no es convergente, pero sin duda hay dos puntos "especiales" para esta sucesión, el 1 y el -1. Quizá el lector se sienta tentado de afirmar que se trata de una sucesión con dos límites, pero nuestra definición de límite no admite esa posibilidad. Vamos a dar una definición que describa esta situación.

Definición 1.80 Un punto x de un espacio topológico X es un *punto adherente* de una sucesión $\{a_n\}_{n=0}^{\infty}$ en X si para cada entorno V de x y cada número natural n existe un número natural $m \geq n$ tal que $a_m \in V$.

Es decir, si la sucesión contiene puntos arbitrariamente lejanos en cualquier entorno de x o, si se prefiere, si la sucesión contiene infinitos puntos en cada entorno de x.

En estos términos, la sucesión $\{(-1)^n\}_{n=0}^{\infty}$ que hemos puesto como ejemplo tiene exactamente dos puntos adherentes, 1 y -1.

Teorema 1.81 *Sea X un espacio 1AN. Un punto $x \in X$ es un punto adherente de una sucesión $\{a_n\}_{n=0}^{\infty}$ si y sólo si ésta contiene una subsucesión que converge a x.*

DEMOSTRACIÓN: Si x es un punto adherente de la sucesión, sea $\{V_n\}_{n=0}^{\infty}$ una base decreciente de entornos de x. Existe un punto $a_{n_0} \in V_0$. Existe un natural $n_1 \geq n_0 + 1$ tal que $a_{n_1} \in V_1$. Existe un natural $n_2 \geq n_1 + 1$ tal que $a_{n_2} \in V_2$. De este modo obtenemos una subsucesión $\{a_{n_k}\}_{k=0}^{\infty}$ tal que cada $a_{n_k} \in V_k$ y, como la sucesión de entornos es decreciente, en realidad V_k contiene todos los términos de la subsucesión posteriores a a_{n_k}, luego la subsucesión que hemos obtenido está finalmente en cada entorno V_k, es decir, converge a x.

Recíprocamente, si hay una subsucesión que converge a x, dicha subsucesión está finalmente en cualquier entorno de x, luego dicho entorno contiene infinitos términos de la sucesión dada. ■

En particular vemos que una sucesión convergente no tiene más punto adherente que su límite.

La continuidad de funciones puede caracterizarse en términos de sucesiones:

Teorema 1.82 *Sea $f : X \longrightarrow Y$ una aplicación entre espacios topológicos y supongamos que X cumple 1AN. Sea $x \in X$. Entonces f es continua en x si y sólo si para cada sucesión $\{a_n\}_{n=0}^{\infty} \subset X$ tal que $\lim\limits_n a_n = x$, se cumple $\lim\limits_n f(a_n) = f(x)$.*

DEMOSTRACIÓN: Supongamos que f es continua en x. Sea V un entorno de $f(x)$. Entonces $f^{-1}[V]$ es un entorno de x y la sucesión $\{a_n\}_{n=0}^{\infty}$ está finalmente en $f^{-1}[V]$, luego $\{f(a_n)\}_{n=0}^{\infty}$ está finalmente en V, o sea, $\lim\limits_n f(a_n) = f(x)$.

Recíprocamente, supongamos que f no es continua en x. Entonces existe un entorno V de $f(x)$ tal que $f^{-1}[V]$ no es entorno de x. Sea $\{V_n\}_{n=0}^{\infty}$ una base decreciente de entornos de x. Para cada natural n, no puede ocurrir que $V_n \subset f^{-1}[V]$, luego existe un punto $a_n \in V_n$ tal que $f(a_n) \notin V$.

Como la base es decreciente, todos los términos posteriores a a_n están en V_n, luego la sucesión $\{a_n\}_{n=0}^{\infty}$ está finalmente en cada V_n, con lo que converge a x. Sin embargo la sucesión $\{f(a_n)\}_{n=0}^{\infty}$ no tiene ningún término en V, luego no converge a $f(x)$. ■

Los puntos adherentes se caracterizan por sucesiones:

Teorema 1.83 *Sea X un espacio topológico que cumpla 1AN. Sea $A \subset X$. Entonces $\overline{A}$ está formado por los límites de las sucesiones convergentes contenidas en A.*

DEMOSTRACIÓN: Si l es el límite de una sucesión contenida en A, entonces todo entorno de l contiene puntos de la sucesión, es decir, puntos de A, luego $l \in \overline{A}$.

Recíprocamente, si $x \in \overline{A}$, tomamos una base decreciente $\{V_n\}_{n=0}^{\infty}$ de entornos abiertos de x. Como $V_n \cap A \neq \varnothing$, existe un $a_n \in V_n \cap A$ y la sucesión $\{a_n\}_{n=0}^{\infty}$ así construida converge a x, y está contenida en A. ■

En particular, un conjunto A es cerrado si y sólo sí el límite de toda sucesión convergente contenida en A, está en A, es decir, si no es posible "salir" de A mediante sucesiones.

Veamos cómo puede aplicarse este hecho: supongamos que una sucesión de números reales cumple $\lim\limits_n a_n = l$ y $a_n \leq 5$ para todo n. Entonces la sucesión está contenida en el intervalo cerrado $]-\infty, 5]$, luego su límite también, es decir, podemos concluir que $l \leq 5$.

Ejercicio: Probar que si dos sucesiones convergentes de números reales cumplen $a_n \leq b_n$ para todo n, entonces $\lim\limits_n a_n \leq \lim\limits_n b_n$.

1.8 Sucesiones y series numéricas

Vamos a estudiar algunos casos concretos de límites de sucesiones en $\mathbb{K}$. Consideremos en primer lugar la sucesión $\{r^n\}_{n=0}^{\infty}$, donde $r \in \mathbb{R}$. Vamos a calcular su límite.

Supongamos primero que $r > 1$. Sea $s = r - 1 > 0$. Entonces $r = 1 + s$ y por el teorema del binomio de Newton

$$r^n = (1+s)^n = 1 + ns + \sum_{k=2}^{n} \binom{n}{k} s^k > ns.$$

Sabemos que $\lim\limits_n ns = +\infty$. Del hecho de que $\{ns\}_{n=0}^{\infty}$ esté finalmente en cada entorno básico de $+\infty$, de la forma $]M, +\infty]$, se sigue claramente que lo mismo le sucede a $\{r^n\}_{n=0}^{\infty}$, luego si $r > 1$ concluimos que $\lim\limits_n r^n = +\infty$.

Si $r = 1$ es obvio que $\lim\limits_n r^n = 1$.

Si $0 < r < 1$ entonces

$$\lim_n r^n = \lim_n \frac{1}{(1/r)^n} = \frac{1}{\infty} = 0.$$

Si $r = 0$ es obvio que $\lim\limits_n r^n = 0$.

En lugar de analizar ahora el caso $r < 0$ consideraremos en general $r \in \mathbb{K}$.

Si $|r| > 1$, entonces $\lim\limits_n |r^n| = \lim\limits_n |r|^n = +\infty$, de donde es fácil deducir a partir de las meras definiciones que $\lim\limits_n r^n = \infty$.

Si $|r| < 1$, entonces $\lim\limits_n |r^n| = 0$, de donde también se sigue que $\lim\limits_n r^n = 0$.

Puede probarse que si $|r| = 1$ el límite no existe salvo en el caso $r = 1$. Por ejemplo, ya hemos visto que $\lim\limits_n (-1)^n$ no existe, pues se trata de la sucesión $1, -1, 1, -1, 1, -1, \ldots$

Definición 1.84 Una sucesión $\{a_n\}_{n=0}^{\infty}$ es *monótona creciente* si $m < n$ implica $a_m \leq a_n$. La sucesión es *estrictamente* creciente si cuando $m < n$ entonces $a_m < a_n$. Igualmente se definen las sucesiones *monótonas decrecientes* y las sucesiones *estrictamente decrecientes.* Una sucesión es *monótona* si cumple cualquiera de estas cuatro propiedades.

El interés de las sucesiones monótonas estriba en que podemos probar que son convergentes sin necesidad de calcular su límite. Esta clase de resultados de convergencia proporcionan la técnica más importante para definir números reales, expresándolos como límites de sucesiones sencillas.

Teorema 1.85 *Toda sucesión monótona creciente converge a su supremo en* $\overline{\mathbb{R}}$, *y toda sucesión monótona decreciente converge a su ínfimo en* $\overline{\mathbb{R}}$.

Demostración: Por supremo de una sucesión $\{a_n\}_{n=0}^{\infty}$ entendemos el supremo del conjunto $\{a_n \mid n \in \mathbb{N}\}$. Sea s este supremo y supongamos que es finito. Entonces un entorno básico de s es de la forma $]s - \epsilon, s + \epsilon[$, para un

$\epsilon > 0$. Como $s - \epsilon$ no es una cota superior de la sucesión, existe un natural m tal que $s - \epsilon < a_m \leq s$ y, por la monotonía, $s - \epsilon < a_n \leq s$ para todo $n \geq m$, es decir, que la sucesión está finalmente en el entorno. Una ligera modificación nos da el mismo resultado si $s = +\infty$ o si la sucesión es decreciente. ■

Prestaremos ahora atención a un tipo especial de sucesiones de particular interés. Se trata de las llamadas series numéricas.

Definición 1.86 Llamaremos *serie* determinada por una sucesión $\{a_n\}_{n=0}^{\infty}$ en $\mathbb{K}$ a la sucesión $\left\{\sum\limits_{n=0}^{k} a_n\right\}_{k=0}^{\infty}$. La representaremos por $\sum\limits_{n=0}^{\infty} a_n$ o, más gráficamente, por

$$a_0 \quad + \quad a_1 \quad + \quad a_2 \quad + \quad a_3 \quad + \quad a_4 \quad + \quad a_5 \quad + \quad \cdots$$

Los términos $\sum\limits_{n=0}^{k} a_n$ se llaman *sumas parciales* de la serie. Si es convergente, su límite se llama *suma* de la serie y se representa también por $\sum\limits_{n=0}^{\infty} a_n$. Esto nunca lleva a confusión.

Así pues, una serie es una suma con infinitos sumandos. Una *serie de términos positivos* es, como su nombre indica, una serie en $\mathbb{R}$ tal que todos los términos a_n son positivos. Vista como sucesión, una serie de términos positivos es estrictamente creciente, luego converge en $\overline{\mathbb{R}}$. De todos modos es costumbre llamar divergente a una serie cuya suma sea $+\infty$.

Una *progresión geométrica* es una sucesión en la que cada término se obtiene del anterior multiplicándolo por un número fijo llamado *razón.* Por lo tanto, si el primer término de una progresión geométrica es a_0 y la razón es r, los términos siguientes serán $a_1 = a_0\, r$, $a_2 = a_0\, r^2$, $a_3 = a_0\, r^3$, $\ldots$ y, en general, $a_n = a_0\, r^n$.

Una *serie geométrica* es una serie asociada a una progresión geométrica, o sea, una serie de la forma $\sum\limits_{n=0}^{\infty} ar^n$.

Una suma parcial es de la forma

$$a(1 + r + r^2 + \cdots + r^n).$$

Es conocida la identidad

$$r^n - 1 = (r - 1)(1 + r + r^2 + \cdots + r^{n-1}),$$

luego si $r \neq 1$ tenemos que

$$a(1 + r + r^2 + \cdots + r^n) = \frac{ar^{n+1} - a}{r - 1} = \frac{a - ar^{n+1}}{1 - r},$$

y de aquí,

$$\sum_{n=1}^{\infty} ar^n = \lim_n \frac{a - ar^{n+1}}{1-r}.$$

Si $|r| > 1$ sabemos que $\lim\limits_n ar^{n+1} = \infty$, de donde se sigue que la serie diverge. Por otro lado, si $|r| < 1$, entonces $\lim\limits_n ar^{n+1} = 0$, luego

$$\sum_{n=1}^{\infty} ar^n = \frac{a}{1-r}.$$

Por ejemplo:

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32} + \cdots = \frac{\frac{1}{2}}{1 - \frac{1}{2}} = 1.$$

Así pues, las series geométricas convergen cuando la razón tiene módulo menor que 1, y su suma es el primer término dividido entre 1 menos la razón.

Ejemplo: expansiones decimales Es conocido que todo número natural n puede expresarse de la forma $n = \sum\limits_{i=1}^{r} a_i 10^i$, donde $a_0, a_1, \ldots, a_r$ son números naturales menores que 10. Además, para $n \neq 0$ la expresión es única si exigimos $a_r \neq 0$. De hecho usamos esta expresión para nombrar a cada número natural, por ejemplo

$$4.275 = 4 \cdot 10^3 + 2 \cdot 10^2 + 7 \cdot 10^1 + 5 \cdot 10^0.$$

Más en general, dado un número natural $k \geq 2$, todo número natural puede expresarse en la forma $\sum\limits_{i=0}^{r} a_i k^i$, donde $a_0, a_1, \ldots, a_r$ son números naturales menores que k.

Vamos a extender esta notación para nombrar a ciertos números racionales. En lo sucesivo, una expresión como ésta: $23,472$ representará al número

$$2 \cdot 10^1 + 3 \cdot 10^0 + 4 \cdot 10^{-1} + 7 \cdot 10^{-2} + 2 \cdot 10^{-3}.$$

Es decir, del mismo modo que el 2 se interpreta multiplicado por 10, el primer número a la derecha de la coma se considerará dividido entre 10, el segundo entre 100, etc. De este modo, los números

$$0 \quad 0,1 \quad 0,2 \quad 0,3 \quad 0,4 \quad 0,5 \quad 0,6 \quad 0,7 \quad 0,8 \quad 0,9 \quad 1$$

dividen al intervalo unidad en 10 partes iguales. Cada una de estas partes se divide a su vez en 10 partes añadiendo una cifra más, por ejemplo:

$$0,3 \quad 0,31 \quad 0,32 \quad 0,33 \quad 0,34 \quad 0,35 \quad 0,36 \quad 0,37 \quad 0,38 \quad 0,39 \quad 0,4$$

es una división del intervalo $[0,3\,,\,0,4]$ en 10 partes iguales.

Por supuesto podemos usar otras bases diferentes de 10. Por ejemplo, en base 10 tenemos que $1/2 = 5/10 = 0,5$. En base dos, en cambio, $1/2 = 0,1$.

Los números racionales expresables de esta forma se llaman números *decimales exactos.*

Tomemos ahora un número natural $k \geq 2$ y sea $\{a_n\}_{n=1}^{\infty}$ una sucesión de números naturales menores que k. Entonces

$$\sum_{n=1}^{r} a_n k^{-n} \leq (k-1)\sum_{n=1}^{r} k^{-n} < (k-1)\sum_{n=1}^{\infty} k^{-n} = (k-1)\frac{\frac{1}{k}}{1-\frac{1}{k}} = 1.$$

Por lo tanto, el supremo de las sumas parciales de $\sum\limits_{n=1}^{\infty} a_n k^{-n}$ es menor o igual que 1, luego $0 \leq \sum\limits_{n=1}^{\infty} a_n k^{-n} \leq 1$.

Esto nos permite extender nuestros desarrollos decimales hasta admitir un número infinito de cifras. Por ejemplo: $532,11111\ldots$ representa al número

$$5 \cdot 10^2 + 3 \cdot 10^1 + 2 \cdot 10^0 + 1 \cdot 10^{-1} + 1 \cdot 10^{-2} + 1 \cdot 10^{-3} + 1 \cdot 10^{-4} + \cdots$$

Como $\sum\limits_{n=1}^{\infty} 10^{-n} = 1/9$, tenemos que $532,11111\ldots = 532 + 1/9$.

Notar que las sucesiones finalmente nulas nos dan los decimales exactos:

$$3,67 = 3,6700000\ldots$$

El interés de usar infinitas cifras decimales es que todo número real positivo es expresable mediante uno de estos desarrollos, es decir, si k es un número natural mayor o igual que 2, todo número real positivo x se puede escribir como

$$x = \sum_{n=0}^{r} b_n k^n + \sum_{n=1}^{\infty} a_n k^{-n},$$

donde los coeficientes son números naturales menores que k.

Vamos a probarlo para $k = 10$ y $x = \sqrt{2}$, aunque el procedimiento es completamente general. La parte $\sum\limits_{n=0}^{r} b_n k^n$ del desarrollo buscado no es sino el desarrollo decimal de la parte entera de x. En nuestro caso, como $1 < 2 < 4$, resulta que $1 < \sqrt{2} < 2$, luego la parte entera es 1.

Dividimos el intervalo $[n, n+1]$ en el que se encuentra x en 10 partes, que en nuestro caso son

$$1 \quad 1,1 \quad 1,2 \quad 1,3 \quad 1,4 \quad 1,5 \quad 1,6 \quad 1,7 \quad 1,8 \quad 1,9 \quad 2$$

Nuestro número x debe estar en uno de los 10 intervalos. En nuestro caso, como

$$\begin{array}{rclrcl} (1,0)^2 &=& 1 & (1,5)^2 &=& 2,25 \\ (1,1)^2 &=& 1,21 & (1,6)^2 &=& 2,56 \\ (1,2)^2 &=& 1,44 & (1,7)^2 &=& 2,89 \\ (1,3)^2 &=& 1,69 & (1,8)^2 &=& 3,24 \\ (1,4)^2 &=& 1,96 & (1,9)^2 &=& 3,61 \end{array}$$

resulta que $(1,4)^2 < 2 < (1,5)^2$, luego $1,4 < \sqrt{2} < 1,5$. Procediendo del mismo modo podemos ir obteniendo las desigualdades siguientes:

$$\begin{array}{rcl} 1 & <\sqrt{2}< & 2 \\ 1,4 & <\sqrt{2}< & 1,5 \\ 1,41 & <\sqrt{2}< & 1,42 \\ 1,414 & <\sqrt{2}< & 1,415 \\ \cdots & \cdots & \cdots \end{array}$$

En general, la n-sima cifra se obtiene como la máxima que hace que el decimal exacto correspondiente sea menor o igual que x. Ahora probemos que la sucesión así obtenida converge a x, es decir, probemos que

$$\sqrt{2} = 1,4142135623730950488016887\ldots$$

No olvidemos que $S = 1,4142135623730950488016887\ldots$ es por definición la suma de una serie cuyas sumas parciales son 1; $1,4$; $1,41$; $1,414\ldots$

Por construcción cada suma parcial es menor que $\sqrt{2}$, luego la suma de la serie, que es el supremo de dichas sumas, cumple $S \leq \sqrt{2}$.

Ahora bien, tenemos que $1 \leq S \leq \sqrt{2} \leq 2$, luego la distancia $|\sqrt{2} - S| \leq |2-1| = 1$. Igualmente $1,4 \leq S \leq \sqrt{2} \leq 1,5$, luego $|\sqrt{2} - S| \leq |1,5-1,4| = 0,1 = 1/10$ y, del mismo modo, $1,41 \leq S \leq \sqrt{2} \leq 1,42$, luego $|\sqrt{2} - S| \leq |1,42 - 1,41| = 0,01 = 1/100$.

En general concluimos que $0 \leq |\sqrt{2} - S| \leq 10^{-n}$ para todo número natural n. Como la sucesión 10^{-n} tiende a 0 concluimos que $0 \leq |\sqrt{2} - S| \leq \epsilon$ para todo $\epsilon > 0$, lo cual sólo es posible si $|\sqrt{2} - S| = 0$, o sea, si $S = \sqrt{2}$.

Esto prueba también que al truncar un número real (es decir, al quedarnos con un número finito de sus decimales) obtenemos una aproximación tanto mejor cuantas más cifras conservemos. Por ejemplo, el número racional $1,4142$ no es $\sqrt{2}$, pero se diferencia de $\sqrt{2}$ en menos de $1/10000$. Su cuadrado no es 2, obviamente, pero es $1,99996164$, que dista de 2 menos de $1/10000$. En muchas ocasiones y a efectos prácticos esto es más que suficiente.

Es importante notar que los desarrollos no son únicos, pues por ejemplo es claro que

$$1 = 0,99999\ldots; \quad 0,1 = 0,099999\ldots; \quad 0,01 = 0,0099999\ldots$$

En general, si una sucesión $\{a_n\}_{n=1}^{\infty}$ es finalmente igual a $k-1$ (digamos a partir de a_n), entonces

$$0,a_1a_2\ldots a_{n-1}(k-1)(k-1)(k-1)\ldots = 0,a_1a_2\ldots a_{n-2}(a_{n-1}+1),$$

pues

$$0,a_1a_2\ldots a_{n-1}(k-1)(k-1)(k-1)\ldots = 0,a_1a_2\ldots a_{n-1} + \sum_{r=n}^{\infty} \frac{k-1}{k^r}$$

$$= 0,a_1a_2\ldots a_{n-1} + \frac{1}{k^{n-1}} = 0,a_1a_2\ldots a_{n-2}(a_{n-1}+1).$$

O sea, toda expresión finalmente igual a $k-1$ admite una expresión exacta sumando 1 a la cifra anterior al primer $k-1$ de la cola constante.

Si no admitimos sucesiones finalmente iguales a $k-1$ el desarrollo es único. En efecto, supongamos que $0,a_1a_2\ldots = 0,b_1b_2\ldots$ Sea $a_n \neq b_n$ la primera cifra diferente. Digamos que $a_n < b_n$. Tenemos que

$$0,a_1a_2\ldots a_{n-1} + \sum_{r=n}^{\infty} b_r k^{-r} = 0,a_1a_2\ldots a_n - 1 + \sum_{r=n}^{\infty} a_r k^{-r},$$

luego

$$\frac{b_n}{k^n} + \sum_{r=n+1}^{\infty} b_r k^{-r} = \frac{a_n}{k^n} + \sum_{r=n+1}^{\infty} a_r k^{-r} < \frac{a_n}{k^n} + \sum_{r=n+1}^{\infty} (k-1)k^{-r} = \frac{a_n}{k^n} + \frac{1}{k^n}.$$

Notar que la desigualdad es estricta porque no todos los a_r son iguales a $k-1$ (no aceptamos desarrollos finalmente iguales a $k-1$). Por lo tanto $b_n < a_n + 1$, o sea, $b_n \leq a_n$, contradicción.

El algoritmo de Euclides para dividir números naturales es válido para calcular el desarrollo decimal de cualquier número racional:

```
1,000000| 7
1 0      0,14284
  30
   20
    60
     30
      2
```

Así pues, $1/7 = 0,1428428428428\ldots$ La razón por la que el algoritmo es válido es que los cálculos que realizamos pueden expresarse de un modo menos práctico, pero conceptualmente más claro, de la forma siguiente:

$$\begin{aligned} \frac{1}{7} &= 0 + \frac{1}{7} = 0 + \frac{1}{10}\cdot\frac{10}{7} = 0 + \frac{1}{10}\cdot\left(1+\frac{3}{7}\right) = 0 + \frac{1}{10}\cdot\frac{1}{100}\cdot\frac{30}{7} \\ &= 0 + \frac{1}{10}\cdot\frac{1}{100}\cdot\left(4+\frac{2}{7}\right) = 0 + \frac{1}{10}\cdot\frac{1}{100}\cdot 4 + \frac{1}{1.000}\cdot\frac{20}{7} = \cdots \end{aligned}$$

Una consecuencia importante es que, como los restos posibles son un número finito, tras un número finito de pasos hemos de obtener un resto ya obtenido previamente, y como cada cifra del cociente depende exclusivamente del último resto, resulta que las cifras se repiten cíclicamente. En el caso de $1/7$ el grupo de cifras que se repite es 428. Para indicar esto se suele usar la notación $1/7 = 0,14\,\overset{\frown}{428}$.

El bloque 428 se suele llamar *período* del número. El bloque de cifras decimales previas al período (que puede no existir) se llama *anteperíodo* (1 en este caso), luego en la expresión decimal de un número racional podemos distinguir la parte entera, el anteperíodo y el período.

Recíprocamente, todo número cuya expresión decimal sea de esta forma es un número racional. Veámoslo con un ejemplo. $r = 37,195\ \overset{\frown}{513}$. Vamos a encontrar una fracción igual a r. El método es general. Multiplicamos por un 1 seguido de tantos ceros como cifras hay en el período más el anteperíodo:

$$(100.000)r = 3.719.513, \overset{\frown}{513}\ .$$

Le restamos r multiplicado por un 1 seguido de tantos ceros como cifras hay en el anteperíodo:

$$(1.000.000)r - (1.000)r = 3.719.513 + 0, \overset{\frown}{513} - 3.719 - 0, \overset{\frown}{513} = 3.715.794.$$

Por consiguiente

$$r = \frac{3.715.794}{999.000}.$$

■

Un problema que surge a menudo es el de determinar si una serie dada an es o no convergente. Lo más elemental que puede decirse al respecto es que para que una serie converja su término general ha de tender a 0. En efecto:

Teorema 1.87 *Sea* $\{a_n\}_{n=0}^{\infty}$ *una sucesión en* $\mathbb{K}$. *Si la serie* $\sum\limits_{n=0}^{\infty} a_n$ *es convergente, entonces* $\lim\limits_n a_n = 0$.

Demostración: Sea $S_k = \sum\limits_{n=0}^{k} a_n$. Que la serie converja a un número L significa por definición que existe $\lim\limits_k S_k = L$. En tal caso también existe $\lim\limits_k S_{k-1} = L$, pues es el límite de una subsucesión, y entonces

$$\lim_k a_k = \lim_k (S_k - S_{k-1}) = L - L = 0.$$

■

Así, si no sumamos cada vez cantidades más pequeñas la serie no puede converger. El recíproco es tentador, pero falso. Basta considerar la serie determinada por la más sencilla de las sucesiones que tienden a 0:

Ejemplo La serie $\sum\limits_{n=1}^{\infty} \frac{1}{n}$ es divergente.

En efecto, observemos que

$$\begin{aligned} S_1 &= 1, \\ S_2 &= 1 + \frac{1}{2}, \\ S_4 &= S_2 + \frac{1}{3} + \frac{1}{4} > S_2 + \frac{2}{4} = 1 + \frac{1}{2} + \frac{1}{2}, \\ S_8 &= S_4 + \frac{1}{5} + \frac{1}{6} + \frac{1}{7} + \frac{1}{8} > S_4 + \frac{4}{8} > 1 + \frac{1}{2} + \frac{1}{2} + \frac{1}{2}. \end{aligned}$$

En general $S_{2^n} > 1 + \frac{n}{2}$, y esta sucesión tiende a $+\infty$, luego las sumas parciales no están acotadas y la serie diverge. ■

Se conocen muchos criterios para determinar el carácter convergente o divergente de una serie. Por ejemplo, el siguiente es aplicable a series de términos positivos.

Teorema 1.88 (Criterio de D'Alembert) *Sea $\{a_n\}_{n=0}^{\infty}$ una sucesión de números reales positivos tal que exista $\lim\limits_{n} \frac{a_{n+1}}{a_n} = L \in \mathbb{R}$. Entonces:*

a) Si $L < 1$ la serie $\{a_n\}_{n=0}^{\infty}$ es convergente.

b) Si $L > 1$ la serie $\{a_n\}_{n=0}^{\infty}$ es divergente.

Demostración: a) Sea $\epsilon > 0$ tal que $L+\epsilon < 1$. Entonces $]-\infty, L+\epsilon[$ es un entorno de L, y por definición de límite existe un natural n_0 tal que si $n \geq n_0$ entonces $\frac{a_{n+1}}{a_n} < L + \epsilon$ o, lo que es lo mismo, $a_{n+1} < a_n(L+\epsilon)$. Así pues,

$$\begin{aligned} a_{n_0+1} &< a_{n_0}(L+\epsilon), \\ a_{n_0+2} &< a_{n_0+1}(L+\epsilon) < a_{n_0}(L+\epsilon)^2, \\ a_{n_0+3} &< a_{n_0+2}(L+\epsilon) < a_{n_0}(L+\epsilon)^3, \ldots \end{aligned}$$

En general, si $n \geq n_0$, se cumple que

$$a_n < a_{n_0}(L+\epsilon)^{n-n_0} = \frac{a_{n_0}}{(L+\epsilon)^{n_0}}(L+\epsilon)^n.$$

De aquí que si $k > n_0$,

$$\sum_{n=0}^{k} a_n \leq \sum_{n=0}^{k} a_n + \frac{a_{n_0}}{(L+\epsilon)^{n_0}} \sum_{n=n_0+1}^{k} (L+\epsilon)^n < \frac{a_{n_0}}{(L+\epsilon)^{n_0}} \sum_{n=n_0+1}^{\infty} (L+\epsilon)^n < +\infty,$$

donde la última serie converge porque es geométrica de razón $L + \epsilon < 1$. La serie $\sum\limits_{n=0}^{\infty} a_n$ es de términos positivos y sus sumas parciales están acotadas, luego converge.

b) Sea ahora $\epsilon > 0$ tal que $1 < L-\epsilon$. Igual que en el apartado anterior existe un natural n_0 tal que si $n \geq n_0$ entonces $a_{n+1} > a_n(L-\epsilon)$, de donde se deduce igualmente que si $n \geq n_0$ entonces

$$a_n > \frac{a_{n_0}}{(L-\epsilon)^{n_0}}(L-e)^n,$$

y por consiguiente, para $k > n_0$,

$$\sum_{n=0}^{k} a_n \geq \sum_{n=0}^{n_0} a_n + \frac{a_{n_0}}{(L-\epsilon)^{n_0}} \sum_{n=n_0+1}^{k} (L-\epsilon)^n,$$

pero ahora la última serie diverge, pues es geométrica de razón mayor que 1, luego sus sumas parciales no están acotadas, y las de la primera serie tampoco. Así pues, ésta es divergente. ■

En el caso de que el límite L exista y valga 1 no es posible asegurar nada, hay casos en los que esto ocurre y la serie converge y casos en los que diverge.

Ejemplo Si $M > 0$, la serie $\sum\limits_{n=0}^{\infty} \frac{M^n}{n!}$ es convergente, pues por el criterio de D'Alembert,

$$L = \lim_n \frac{M^{n+1}}{(n+1)!} : \frac{M^n}{n!} = \lim_n \frac{M}{n+1} = 0 < 1.$$

Incidentalmente, esto prueba también que $\lim\limits_n \frac{M^n}{n!} = 0$.

Notar que hemos determinado el carácter de la serie, pero no hemos dicho nada sobre el cálculo efectivo de su límite. La razón es que no hay nada que decir. Por ejemplo, en el caso más simple, $M = 1$, el número

$$e = \sum_{n=0}^{\infty} \frac{1}{n!}$$

es un número "nuevo", en el sentido de que no es racional, ni la raíz cuadrada de un número racional, ni en general expresable en términos de otros números ya conocidos. Más adelante demostraremos que de hecho se trata de un número trascendente sobre $\mathbb{Q}$. El único sentido en que podemos "calcularlo" es en el de obtener aproximaciones racionales sumando términos de la serie. El resultado es

$$e = 2,7182818284590452353602874\ldots$$

■

Para acabar demostraremos un criterio válido para las llamadas series alternadas, es decir, para series de números reales en las que los términos sucesivos tienen signos opuestos:

Teorema 1.89 (Criterio de Leibniz) *Sea $\{a_n\}_{n=0}^{\infty}$ una sucesión de números reales positivos decreciente y convergente a 0. Entonces la serie $\sum\limits_{n=0}^{\infty} (-1)^n a_n$ es convergente.*

DEMOSTRACIÓN: Consideremos primero las sumas parciales pares. Por ejemplo:

$$S_6 = (a_0 - a_1) + (a_2 - a_3) + (a_4 - a_5) + a_6.$$

Teniendo en cuenta que la sucesión es decreciente, los sumandos así agrupados son todos mayores o iguales que 0, luego en general $S_{2n} \geq 0$.

Por otra parte, $S_8 = S_6 + (-a_7 + a8) \leq S_6$, luego la sucesión $\{S_{2n}\}_{n=0}^{\infty}$ es monótona decreciente y acotada inferiormente por 0. Por lo tanto converge a un número L.

Ahora, $S_{2n+1} = S_{2n} + a_{2n+1}$, luego existe $\lim\limits_n S_{2n+1} = L + 0$.

Es fácil comprobar que si las dos subsucesiones $\{S_{2n}\}_{n=0}^{\infty}$ y $\{S_{2n+1}\}_{n=0}^{\infty}$ convergen a un mismo número L, entonces toda la sucesión $\{S_n\}_{n=0}^{\infty}$ converge a L, es decir, la serie converge. ■

Por ejemplo, la serie $\sum\limits_{n=1}^{\infty} (-1)^{n+1} \frac{1}{n}$ es convergente. De nuevo no tenemos más medio para calcular su límite que aproximarlo por una suma parcial. En

realidad, cuando hacemos esto necesitamos saber cuál es el error cometido, para determinar el número de cifras decimales correctas. Por ejemplo, las primeras sumas parciales de esta serie son:

1	1	11	0, 73654	21	0, 71639	31	0, 70901
2	0, 50000	12	0, 65321	22	0, 67093	32	0, 67776
3	0, 83333	13	0, 73013	23	0, 71441	33	0, 70806
4	0, 58333	14	0, 65870	24	0, 67274	34	0, 67865
5	0, 78333	15	0, 72537	25	0, 71274	35	0, 70722
6	0, 61666	16	0, 66287	26	0, 67428	36	0, 67945
7	0, 75952	17	0, 72169	27	0, 71132	37	0, 70647
8	0, 63452	18	0, 66613	28	0, 67560	38	0, 68016
9	0, 74563	19	0, 71877	29	0, 71009	39	0, 70580
10	0, 64563	20	0, 66877	30	0, 67675	40	0, 68080

El límite vale $0,6931471805599453094172321\ldots$ Por lo tanto, si al calcular la décima suma afirmáramos que el límite vale $0,6456\ldots$ estaríamos cometiendo un grave error. En realidad sólo la primera cifra decimal es exacta (se dice que un número decimal finito aproxima a otro con n cifras exactas si las n primeras cifras de ambos números coinciden). En general, al aproximar una serie por una suma parcial, o al aproximar cualquier límite de una sucesión por uno de sus términos, no sabemos cuál es el error cometido, ni en particular cuántas de las cifras de la aproximación son exactas. Sin embargo, en el caso de las series alternadas es fácil saberlo pues, según hemos visto, las sumas pares son decrecientes (siempre están sobre el límite) y las impares son crecientes (siempre están bajo el límite), luego las últimas sumas de la tabla nos dicen que el límite se encuentra entre 0,68 y 0,71 y por lo tanto no sabemos ninguna de sus cifras con exactitud (salvo el 0). Yendo más lejos tenemos:

$$S_{1000} = 0,69264 \quad \text{y} \quad S_{1001} = 0,69364,$$

lo que nos permite afirmar que el límite está entre $0,692$ y $0,694$, es decir, es de la forma $0,69\ldots$ y ya tenemos dos cifras exactas.

En la práctica ignoraremos el problema técnico de determinar las cifras exactas que nos proporciona una aproximación dada, y consideraremos que un número es "conocido" si tenemos una sucesión que converge a él. Todas las aproximaciones que daremos (calculadas con ordenador con técnicas de análisis numérico) tendrán todas sus cifras exactas.

Capítulo II

Compacidad, conexión y completitud

Finalmente tenemos los suficientes elementos de topología como para que ésta se convierta en una herramienta eficaz. En los temas anteriores apenas hemos extraído las consecuencias más elementales de las definiciones de topología, continuidad, etc. Los resultados que veremos ahora van mucho más lejos y dan una primera muestra de las posibilidades de las técnicas topológicas.

2.1 Espacios compactos

La compacidad es en topología una propiedad similar a la "dimensión finita" en álgebra lineal. Los espacios compactos no son necesariamente finitos, pero se comportan en muchos aspectos como si lo fueran. Por ejemplo, es obvio que en un espacio finito toda sucesión ha de tomar infinitas veces un mismo valor, luego toda sucesión contiene una subsucesión constante, en particular convergente. Lo mismo ocurre en $\overline{\mathbb{R}}$, como se desprende del teorema siguiente.

Teorema 2.1 *Toda sucesión en un conjunto totalmente ordenado contiene una subsucesión monótona.*

DEMOSTRACIÓN: Sea $\{a_n\}_{n=0}^{\infty}$ una sucesión en un conjunto totalmente ordenado. Sea A el conjunto de las imágenes de la sucesión. Si A es finito es obvio que A tiene una subsucesión constante, luego monótona. Supongamos que A es infinito.

Si todo subconjunto no vacío de A tiene mínimo podemos tomar x_0 igual al mínimo de A, luego x_1 igual al mínimo de $A \setminus \{x_0\}$, luego x_2 igual al mínimo de $A \setminus \{x_0, x_1\}$, y así obtenemos puntos $x_0 < x_1 < x_2 < \ldots$, es decir, obtenemos un subconjunto de A sin máximo.

Así pues, o bien existe un subconjunto de A sin mínimo o bien existe un subconjunto de A sin máximo. Los dos casos se tratan igual. Supongamos que hay un subconjunto de A sin mínimo. Llamémoslo B.

Sea a_{n_0} un elemento cualquiera de B. Como B no tiene mínimo contiene infinitos términos de la sucesión bajo a_{n_0}, pero sólo un número finito de ellos tienen índice anterior a n_0, luego existe un a_{n_1} en B tal que $a_{n_1} < a_{n_0}$ y $n_0 < n_1$. Podemos repetir indefinidamente este proceso y obtener una subsucesión

$$a_{n_0} > a_{n_1} > a_{n_2} > a_{n_3} > a_{n_4} > a_{n_5} > \ldots$$

monótona decreciente. ■

Como en $\overline{\mathbb{R}}$ toda sucesión monótona converge a su supremo o a su ínfimo, esto prueba que en este espacio toda sucesión contiene una subsucesión convergente, al igual que ocurre en los espacios finitos. Cualquier intervalo $[a, b]$ es homeomorfo a $\overline{\mathbb{R}}$, luego también cumple esto mismo.

Por otro lado esto es falso en $\mathbb{R}$. La sucesión de los números naturales no contiene ninguna subsucesión convergente (ya que cualquier subsucesión converge a $+\infty$ en $\overline{\mathbb{R}}$, luego no converge en $\mathbb{R}$).

El espacio $\overline{\mathbb{R}}$ y los intervalos $[a, b]$ son ejemplos de espacios compactos. La propiedad de las subsucesiones convergentes caracteriza la compacidad en espacios métricos, pero para el caso general necesitamos otra definición más elaborada.

Definición 2.2 Sea X un espacio topológico. Un *cubrimiento abierto* de X es una familia $\{A_i\}_{i\in I}$ de abiertos de X tal que $X = \bigcup\limits_{i\in I} A_i$.

Un *subcubrimiento* del cubrimiento dado es un cubrimiento formado por parte de los abiertos del primero.

Un espacio de Hausdorff K es *compacto* si de todo cubrimiento abierto de K se puede extraer un subcubrimiento finito.

Es obvio que si X es un espacio finito, de todo cubrimiento abierto se puede extraer un subcubrimiento finito. Basta tomar un abierto que contenga a cada uno de los puntos del espacio. Así pues, todo espacio de Hausdorff finito es compacto.

Observar que si $\mathcal{B}$ es una base de un espacio de Hausdorff K, se cumple que K es compacto si y sólo si todo cubrimiento de K por abiertos básicos admite un subcubrimiento finito. En efecto, si $\{A_i\}_{i\in I}$ es un cubrimiento arbitrario, para cada punto $x \in K$ existe un $i_x \in I$ tal que $x \in A_{i_x}$ y existe un $B_x \in \mathcal{B}$ tal que $x \in B_x \subset A_{i_x}$. Entonces $\{B_x\}_{x\in K}$ es un cubrimiento de K formado por abiertos básicos y tiene un subcubrimiento finito

$$K = B_{x_1} \cup \cdots \cup B_{x_n} \subset A_{i_{x_1}} \cup \cdots \cup A_{i_{x_n}} \subset K.$$

Una familia de abiertos forma un cubrimiento si y sólo si la familia de sus complementarios es una familia de cerrados con intersección vacía. Por ello la compacidad puede caracterizarse así en términos de familias de cerrados:

Un espacio de Hausdorff K es compacto si y sólo si toda familia de cerrados $\{C_i\}_{i\in I}$ con la propiedad de que cualquier intersección finita de ellos no es vacía, tiene intersección total no vacía.

La propiedad de que las intersecciones finitas sean no vacías se llama *propiedad de la intersección finita.* Por lo tanto:

Teorema 2.3 *Un espacio de Hausdorff K es compacto si y sólo si toda familia de cerrados de K con la propiedad de la intersección finita tiene intersección no vacía.*

A menudo nos encontraremos con espacios que no son compactos pero tienen subespacios compactos. Por ello resulta útil caracterizar la compacidad de un subespacio en términos de la topología de todo el espacio y no de la topología relativa. Concretamente:

Teorema 2.4 *Sea X un espacio de Hausdorff y K un subespacio de X. Entonces K es compacto si y sólo si para toda familia $\{A_i\}_{i\in I}$ de abiertos (básicos) de X tal que $K \subset \bigcup_{i\in I} A_i$ se puede extraer una subfamilia finita que cumpla lo mismo.*

Demostración: Supongamos que K es compacto. Entonces $\{A_i \cap K\}_{i\in I}$ es claramente un cubrimiento abierto de K, del que podemos extraer un subcubrimiento finito de modo que

$$K = (A_{i_1} \cap K) \cup \cdots \cup (A_{i_n} \cap K),$$

luego $K \subset A_{i_1} \cup \cdots \cup A_{i_n}$.

Recíprocamente, si K cumple esta propiedad y $\{A_i\}_{i\in I}$ es un cubrimiento abierto de K, entonces para cada i existe un abierto B_i de X tal que $A_i = B_i \cap K$. Consecuentemente $K = \bigcup_{i\in I} A_i \subset \bigcup_{i\in I} B_i$, luego por hipótesis podemos tomar un número finito de conjuntos de modo que $K \subset B_{i_1} \cup \cdots \cup B_{i_n}$, luego $K = (B_{i_1} \cap K) \cup \cdots \cup (B_{i_n} \cap K) = A_{i_1} \cup \cdots \cup A_{i_n}$. Así pues, K es compacto. ■

Si la unión de una familia de abiertos de un espacio X contiene a un subespacio K, diremos que forma un *cubrimiento abierto* de K en X. Así pues, un subespacio K de X es compacto si y sólo si de todo cubrimiento abierto de K en X puede extraerse un subcubrimiento finito (en X también).

Aquí estamos considerando la topología de X, pero deberemos tener siempre presente que la compacidad es una propiedad absoluta, y depende exclusivamente de la topología del propio espacio K.

Los teoremas siguientes muestran la anunciada similitud entre los espacios compactos y los espacios finitos. Por lo pronto, todo espacio finito es cerrado. El análogo con compactos es el siguiente:

Teorema 2.5 *Se cumplen las propiedades siguientes:*

a) Si X es un espacio de Hausdorff y $K \subset X$ es compacto, entonces K es cerrado en X.

b) Si K es un compacto y $C \subset K$ es un cerrado, entonces C es compacto.

c) Si M es un espacio métrico y $K \subset M$ es compacto, entonces K está acotado.

DEMOSTRACIÓN: a) El argumento es el mismo que emplearíamos si K fuera finito. Veamos que $X \setminus K$ es abierto. Sea $x \in X \setminus K$. Para cada punto $u \in K$ existen abiertos disjuntos A_u y B_u tales que $u \in A_u$ y $x \in B_u$. Si K fuera finito bastaría tomar ahora la intersección de los B_u y tendríamos un entorno de x contenido en $X \setminus K$. Ahora aplicamos la compacidad de K. Los conjuntos A_u forman un cubrimiento abierto de K, luego existe un subcubrimiento finito: $K \subset A_{u_1} \cup \cdots \cup A_{u_n}$. Ahora, $\bigcap\limits_{i=1}^{n} B_{u_i}$ es un entorno de x que no corta a K, luego $X \setminus K$ es un entorno de x.

b) Si $\{A_i\}_{i \in I}$ es un cubrimiento abierto de C, entonces $\{A_i\}_{i \in I} \cup \{K \setminus C\} es$ un cubrimiento abierto de K, luego existe un subcubrimiento finito

$$K = A_{i_1} \cup \cdots \cup A_{i_n} \cup (K \setminus C).$$

Claramente entonces $C \subset A_{i_1} \cup \cdots \cup A_{i_n}$, luego C es compacto.

c) Sea $x \in M$ un punto cualquiera. Para cada $u \in K$, sea $r_u = d(x, u) + 1$. Obviamente $K \subset \bigcup\limits_{u \in K} B_{r_u}(x)$. Por compacidad podemos extraer un subcubrimiento finito de modo que $K \subset B_{r_{u_1}}(x) \cup \cdots \cup B_{r_{u_n}}(x)$. Las bolas son conjuntos acotados, una unión finita de acotados es acotada y todo subconjunto de un acotado está acotado. Por tanto K está acotado. ■

Teorema 2.6 *Si K es un espacio compacto, toda sucesión en K posee un punto adherente. Por tanto si además K cumple 1AN, toda sucesión en K tiene una subsucesión convergente.*

DEMOSTRACIÓN: Sea $\{a_n\}_{n=0}^{\infty}$ una sucesión en K. Sea $A_n = \{a_m \mid m \geq n\}$. Obviamente

$$A_0 \supset A_1 \supset A_2 \supset A_3 \supset A_4 \supset \ldots \quad ,$$

luego también

$$\overline{A}_0 \supset \overline{A}_1 \supset \overline{A}_2 \supset \overline{A}_3 \supset \overline{A}_4 \supset \ldots \quad ,$$

y así tenemos una familia de cerrados con la propiedad de la intersección finita. Por compacidad existe un punto $x \in \bigcap\limits_{i=0}^{\infty} \overline{A}_i$. Obviamente x es un punto adherente de la sucesión, pues si n es un natural y U es un entorno de x, entonces $x \in \overline{A}_n$, luego $U \cap A_n \neq \varnothing$, es decir, existe un $m \geq n$ tal que $a_m \in U$. ■

Como habíamos anunciado, esta propiedad caracteriza a los espacios métricos compactos.

Teorema 2.7 *Un espacio métrico M es compacto si y sólo si toda sucesión en M tiene una subsucesión convergente.*

DEMOSTRACIÓN: Supongamos que M no fuera compacto. Entonces existiría un cubrimiento abierto $M = \bigcup_{i \in I} A_i$ que no admite subcubrimientos finitos.

Sea $\epsilon > 0$ y $x_0 \in M$. Si $B_\epsilon(x) \neq M$, existe un punto $x_1 \in M$ tal que $d(x_1, x_0) \geq \epsilon$. Si $B_\epsilon(x_0) \cup B_\epsilon(x_1) \neq M$, existe un punto $x_2 \in M$ tal que $d(x_2, x_0) \geq \epsilon$, $d(x_2, x_1) \geq \epsilon$.

Si M no pudiera cubrirse por un número finito de bolas de radio ϵ, podríamos construir una sucesión $\{x_n\}_{n=0}^\infty$ con la propiedad de que $d(x_i, x_j) \geq \epsilon$ para todos los naturales i, j. Es claro que tal sucesión no puede tener subsucesiones convergentes, pues una bola de centro el límite y radio $\epsilon/2$ debería contener infinitos términos de la sucesión, que distarían entre sí menos de ϵ.

Concluimos que para todo $\epsilon > 0$ existen puntos $x_0, \ldots, x_n \in M$ de modo que $M = B_\epsilon(x_0) \cup \cdots \cup B_\epsilon(x_n)$.

Lo aplicamos a $\epsilon = 1$ y obtenemos tales bolas. Si todas ellas pudieran cubrirse con un número finito de abiertos A_i también M podría, luego al menos una de ellas, digamos $B_1(x_0)$ no es cubrible por un número finito de abiertos del cubrimiento.

Igualmente, con $\epsilon = 1/2$ obtenemos una bola $B_{1/2}(x_1)$ no cubrible por un número finito de abiertos del cubrimiento. En general obtenemos una sucesión de bolas $B_{1/(n+1)}(x_n)$ con esta propiedad.

Sea x un punto adherente de la sucesión $\{x_n\}_{n=0}^\infty$. Sea $i \in I$ tal que $x \in A_i$. Como A_i es un abierto existe un número natural k tal que $B_{2/(k+1)}(x) \subset A_i$. Sea $n > k$ tal que $d(x_n, x) < 1/(k+1)$. Entonces $B_{1/(n+1)}(x_n) \subset B_{2/(k+1)}(x) \subset A_i$, en contradicción con que $B_{1/(n+1)}(x_n)$ no era cubrible con un número finito de abiertos del cubrimiento. ■

Según lo visto al comienzo de la sección, $\overline{\mathbb{R}}$ es compacto. Además:

Teorema 2.8 *Un subconjunto de $\mathbb{R}$ es compacto si y sólo si es cerrado y acotado.*

DEMOSTRACIÓN: Por el teorema 2.5, todo compacto en $\mathbb{R}$ ha de ser cerrado y acotado. Si C es un conjunto cerrado y acotado, toda sucesión en C tiene una subsucesión convergente en $\overline{\mathbb{R}}$. Como C es acotado su límite estará en $\mathbb{R}$ y como C es cerrado, su límite estará en C, luego toda sucesión en C tiene una subsucesión convergente en C. Por el teorema anterior C es compacto. ■

También se puede probar el teorema anterior viendo que los cerrados y acotados de $\mathbb{R}$ son precisamente los subconjuntos cerrados de $\overline{\mathbb{R}}$ (o, si se prefiere, esto es consecuencia inmediata del teorema anterior).

Ejemplo Vamos a usar la compacidad de $[0, 1]$ para calcular el cardinal de $\mathbb{R}$. Si $I = [a, b]$ es un intervalo cerrado, llamaremos

$$I_0 = \left[a, a + \frac{b-a}{3}\right], \quad I_1 = \left[a + 2\,\frac{b-a}{3}, b\right],$$

que son dos intervalos cerrados disjuntos contenidos en I.

De este modo, partiendo de $I = [0, 1]$ podemos formar los intervalos I_0, I_1, I_{00}, I_{01}, I_{10}, I_{11}, etc. Más exactamente, para cada aplicación $s : \mathbb{N} \longrightarrow \{0, 1\}$ y cada $n \in \mathbb{N}$, si llamamos $s|_n = s_0 \ldots s_{n-1}$, tenemos definidos los intervalos $I_{s|_n}$ (entendiendo que $I_{s|_0} = I$), y es claro que forman una sucesión decreciente, es decir, si $m \leq n$ entonces $I_{s|_n} \subset I_{s|_m}$. Por lo tanto, $\{I_{s|_n}\}_{n=0}^{\infty}$ es una familia de cerrados en I con la propiedad de la intersección finita. Por compacidad tenemos que $I_s = \bigcap\limits_{n=0}^{\infty} I_{s|_n} \neq \varnothing$. Más aún, es claro que la longitud (el diámetro) de $I_{s|_n}$ es $1/3^n$, y como $I_s \subset I_{s|_n}$, necesariamente el diámetro de I_s ha de ser menor o igual que $1/3^n$ para todo n, es decir, ha de ser 0. Esto implica que I_s contiene un único punto. Digamos $I_s = \{x_s\}$.

También es claro que si $s \neq t$ entonces $x_s \neq x_t$. En efecto, si n es el primer natural tal que $s|_{n+1} \neq t_{n+1}$, entonces $I_{s|_n} = I_{t|_n}$ y los intervalos $I_{s|_{n+1}}$, $I_{t|_{n+1}}$ son subconjuntos disjuntos. Como contienen a x_s y x_t respectivamente, éstos han de ser puntos distintos.

Esto prueba que $|I| \geq |2^{\mathbb{N}}| = 2^{\aleph_0}$. Por otra parte, $|\mathbb{R}| \leq 2^{\aleph_0}$, pues si a cada $r \in \mathbb{R}$ le asignamos el conjunto de los números racionales menores que r obtenemos una aplicación inyectiva de $\mathbb{R}$ en las partes de $\mathbb{Q}$. Por consiguiente tenemos que $|I| = |\mathbb{R}| = 2^{\aleph_0}$. ■

Teorema 2.9 (Teorema de Tychonoff) *El producto de espacios compactos es compacto.*

Demostración: Sea $K = \prod\limits_{i \in I} K_i$ un producto de espacios compactos. Tomemos una familia $\mathcal{B}$ de cerrados en K con la propiedad de la intersección finita. Hemos de probar que su intersección es no vacía. El conjunto de todas las familias de subconjuntos no vacíos de K (no necesariamente cerrados) que contienen a $\mathcal{B}$ y tienen la propiedad de la intersección finita, parcialmente ordenado por la inclusión, satisface las hipótesis del lema de Zorn, lo que nos permite tomar una familia maximal $\mathcal{U}$. Entonces $\bigcap\limits_{U \in \mathcal{U}} \overline{U} \subset \bigcap\limits_{B \in \mathcal{B}} B$, luego basta probar que la primera intersección es no vacía.

En primer lugar observamos que si un conjunto $A \subset K$ corta a todos los elementos de $\mathcal{U}$ entonces está en $\mathcal{U}$, pues en caso contrario $\mathcal{U} \cup \{A\}$ contradiría la maximalidad de $\mathcal{U}$.

Sea $p_i : K \longrightarrow K_i$ la proyección en el factor i-ésimo. Es fácil ver que la familia $\{p_i[U] \mid U \in \mathcal{U}\}$ tiene la propiedad de la intersección finita luego, por la compacidad de K_i, existe un punto $x_i \in K_i$ tal que $x_i \in \overline{p_i[U]}$ para todo $U \in \mathcal{U}$. Estos puntos determinan un punto $x \in K$. Basta probar que $x \in \bigcap\limits_{U \in \mathcal{U}} \overline{U}$.

Fijemos un entorno básico de x, de la forma $A = \bigcap\limits_{i \in F} p_i^{-1}[G_i]$, donde $F \subset I$ es finito y G_i es abierto en K_i. Para cada $U \in \mathcal{U}$ tenemos que $x_i \in \overline{p_i[U]}$, luego $G_i \cap p_i[U] \neq \varnothing$, luego $p_i^{-1}[G_i] \cap U \neq \varnothing$. Como esto es cierto para todo $U \in \mathcal{U}$, según hemos observado antes podemos concluir que $p_i^{-1}[G_i] \in \mathcal{U}$, para todo $i \in F$. Como $\mathcal{U}$ tiene la propiedad de la intersección finita, $A \in \mathcal{U}$. De aquí se sigue que A corta a todo $U \in \mathcal{U}$ y, como A es un entorno básico de x, esto implica que $x \in \overline{U}$ para todo $U \in \mathcal{U}$. ■

Ahora podemos probar:

Teorema 2.10 *Un subconjunto de $\mathbb{K}^n$, es compacto si y sólo si es cerrado y acotado.*

DEMOSTRACIÓN: La acotación depende en principio de la distancia que consideremos. Hemos de entender que se trata de la inducida por cualquiera de las tres normas definidas en el capítulo I. Por el teorema 1.5, todas tienen los mismos acotados. Trabajaremos concretamente con

$$\|x\|_\infty = \text{máx}\big\{|x_i| \bigm| i = 1, \ldots, n\big\}.$$

Como $\mathbb{C}^n$ es homeomorfo a $\mathbb{R}^{2n}$ (con los mismos conjuntos acotados), basta probar el teorema para $\mathbb{R}^n$. Ya hemos visto que un compacto ha de ser cerrado y acotado. Supongamos que K es un subconjunto de $\mathbb{R}^n$ cerrado y acotado. Esto significa que existe un $M > 0$ tal que para todo punto $x \in C$ se cumple $\|x\| \leq M$, lo que significa que si $x \in C$, cada $x_i \in [-M, M]$ o, de otro modo, que $C \subset [-M, M]^n$.

Pero por el teorema anterior $[-M, M]^n$ es compacto y C es cerrado en él, luego C también es compacto. ■

Los espacios que nos van a interesar son fundamentalmente los subconjuntos de $\mathbb{R}^n$. Vemos, pues, que la compacidad es muy sencilla de reconocer. En particular las bolas cerradas y las esferas son compactos. El espacio $\mathbb{C}^\infty$ es homeomorfo a una esfera, luego es compacto. Lo mismo ocurre con $\mathbb{R}^\infty$, que es homeomorfo a una circunferencia.

Una de las propiedades más importantes de la compacidad es que se conserva por aplicaciones continuas (compárese con el hecho de que la imagen (continua) de un conjunto finito es finita).

Teorema 2.11 *La imagen de un espacio compacto por una aplicación continua es de nuevo un espacio compacto (supuesto que sea un espacio de Hausdorff).*

DEMOSTRACIÓN: Sea $f : K \longrightarrow X$ continua y suprayectiva. Supongamos que K es compacto y que X es un espacio de Hausdorff. Si $\{A_i\}_{i\in I}$ es un cubrimiento abierto de X, entonces $\big\{f^{-1}[A_i]\big\}_{i\in I}$ es un cubrimiento abierto de K, luego admite un subcubrimiento finito $K = f^{-1}[A_{i_1}] \cup \cdots \cup f^{-1}[A_{i_n}]$. Entonces $X = A_{i_1} \cup \cdots \cup A_{i_n}$. ■

Este hecho tiene muchas consecuencias.

Teorema 2.12 *Si $f : K \longrightarrow X$ es biyectiva y continua, K es compacto y X es un espacio de Hausdorff, entonces f es un homeomorfismo.*

DEMOSTRACIÓN: Basta probar que la inversa en continua, o sea, que transforma cerrados de K en cerrados de X, o equivalentemente, que si C es cerrado en K, entonces $f[C]$ es cerrado en X, pero es que C es compacto, luego $f[C]$ también lo es, luego es cerrado. ■

Ejemplo Una circunferencia es homeomorfa a un cuadrado.

En efecto, si C es un cuadrado de centro $(0,0)$ en $\mathbb{R}^2$, es claro que la aplicación de C en la circunferencia unidad dada por $x \mapsto x/\|x\|$ es biyectiva y continua y, como C es compacto, es un homeomorfismo. ■

***Ejemplo** Los espacios proyectivos son compactos.

En efecto, basta probar que $\mathrm{P}^n(\mathbb{K})$ es compacto, pero la restricción de la proyección a la esfera unidad de $\mathbb{K}^{n+1}$ es continua y suprayectiva y la esfera es compacta. ■

Otro hecho obvio es que toda aplicación continua de un compacto a un espacio métrico está acotada. Para las funciones reales podemos decir más:

Teorema 2.13 *Si $f : K \longrightarrow \mathbb{R}$ es continua y K es un compacto no vacío, existen $u, v \in K$ tales que para todo $x \in K$, se cumple $f(u) \leq f(x) \leq f(v)$. Es decir, que f alcanza un valor mínimo y un valor máximo.*

Demostración: Sea $C = f[K]$. Entonces C es cerrado y acotado. Sean m y M su ínfimo y su supremo, respectivamente. Así para todo $x \in K$ se cumple que $m \leq f(x) \leq M$. Sólo falta probar que m y M son imágenes de puntos de K, o sea, que $m, M \in C$. Veámoslo para M.

Si $\epsilon > 0$, entonces $M - \epsilon$ no es una cota superior de C, luego existe un punto $y \in C$ de modo que $M - \epsilon < y$, es decir, que $]M-\epsilon, M+\epsilon[\, \cap C \neq \varnothing$. Esto significa que todo entorno (básico) de M corta a C, o sea, $M \in \overline{C} = C$. ■

Observar que este resultado es falso sin compacidad. Por ejemplo la función $f :]0,1[\longrightarrow \mathbb{R}$ dada por $f(x) = x$ no tiene máximo ni mínimo. Veamos una aplicación del teorema anterior:

Teorema 2.14 *Todas las normas en $\mathbb{K}^n$ inducen la misma topología. Además los subconjuntos acotados son los mismos para todas ellas.*

Demostración: Sea $\|\ \| : \mathbb{K}^n \longrightarrow [0, +\infty[$ una norma cualquiera. Vamos a ver que induce la misma topología que la dada por $\|x\|_1 = \sum_{i=1}^n |x_i|$ y que los acotados son los mismos para ambas. Esto probará el teorema.

Sea $\{e_1, \ldots, e_n\}$ la base canónica de $\mathbb{K}^n$. Si $x \in \mathbb{K}^n$ se puede expresar como $x = \sum_{i=1}^n x_i e_i$, luego

$$\|x\| \leq \sum_{i=1}^n |x_i|\, \|e_i\| \leq \sum_{i=1}^n \|x\|_1 \|e_i\| = \|x\|_1 \sum_{i=1}^n \|e_i\| = K\|x\|_1.$$

Por lo tanto, si $x, y \in \mathbb{K}^n$, se cumple $\big|\|x\| - \|y\|\big| \leq \|x - y\| \leq K\|x - y\|_1$, o sea que la aplicación $\|\ \|$ tiene la propiedad de Lipschitz, luego es continua respecto a la topología inducida por $\|\ \|_1$.

Sea $S = \{x \in \mathbb{K}^n \mid \|x\|_1 = 1\}$. Se trata de una esfera, luego de un conjunto compacto para lo topología usual de $\mathbb{K}^n$. Por lo tanto la aplicación $\|\ \|$ alcanza

su máximo y su mínimo en S y así, como no se anula, existen $0 < m \leq M$ en $\mathbb{R}$ tales que para todo $x \in S$ se tiene $m \leq \|x\| \leq M$.

Si $x \in \mathbb{K}^n \setminus \{0\}$, entonces $x/\|x\|_1 \in S$, luego $m \leq \big\|x/\|x\|_1\big\| \leq M$, es decir,

$$m\|x\|_1 \leq \|x\| \leq M\|x\|_1,$$

o también,

$$\|x\|_1 \leq \frac{1}{m}\,\|x\|, \quad \|x\| \leq M\|x\|_1.$$

Como 0 cumple esto trivialmente, en realidad vale para todo $x \in \mathbb{K}^n$. De aquí se sigue que toda bola de centro x con respecto a una norma contiene a otra respecto a la otra norma, luego los acotados son los mismos. Además todo entorno de un punto para una norma lo es para la otra norma, luego las topologías inducidas son las mismas. ■

Así pues, podemos hablar de acotados en $\mathbb{K}^n$ sin precisar la norma a la que nos referimos. Sin embargo no hay que olvidar que los acotados pueden variar si consideramos métricas que no provengan de normas. Por ejemplo la distancia

$$d(x, y) = \text{mín}\{\|x - y\|, 1\}.$$

También es claro que el teorema anterior es válido de hecho para cualquier $\mathbb{K}$-espacio vectorial de dimensión finita.

2.2 Espacios conexos

Pensemos en los espacios siguientes: $[0, 1]$ y $[0, 1] \cup [2, 3]$. Hay una diferencia esencial entre ellos, y es que el primero está formado por "una sola pieza" mientras que el segundo consta de "dos piezas". La diferencia no es conjuntista, pues también podemos dividir $[0, 1] = [0, 1/2] \cup \,]1/2, 1]$, pero esto no son dos piezas en el mismo sentido que en el caso de $[0, 1] \cup [2, 3]$. La diferencia es que los intervalos $[0, 1/2]$ y $]1/2, 1]$ están "pegados" mientras que los intervalos $[0, 1]$ y $[2, 3]$ están "separados". Con más precisión, el punto $1/2$ está sólo en uno de los intervalos, el $[0, 1/2]$, pero aunque no está en el otro, está pegado a él, en el sentido de que está en su clausura.

En general, si un espacio X se expresa como $X = U \cup V$, donde U y V son disjuntos y no vacíos, podemos decir que U y V son dos "piezas" en el sentido que estamos considerando si U no contiene puntos de la clausura de V y viceversa. Ahora bien, cualquier punto de $\overline{V}$ que no estuviera en V debería estar en U, luego la condición equivale a que $V = \overline{V}$ y $U = \overline{U}$, o sea, a que U y V sean cerrados. Por otra parte, dado que U y V son complementarios, es lo mismo decir que son cerrados o que son abiertos. Con ello llegamos a la definición de conexión:

Definición 2.15 Un espacio topológico X es *disconexo* si existen subconjuntos abiertos U y V en X tales que $X = U \cup V$, $U \cap V = \varnothing$ y $U \neq \varnothing \neq V$. En caso contrario X es *conexo*.

Según hemos dicho, es indistinto exigir que U y V sean abiertos como que sean cerrados, pues de hecho si cumplen esto son a la vez abiertos y cerrados. Por lo tanto, un espacio X es conexo si y sólo si sus únicos subconjuntos que son a la vez abiertos y cerrados son X y $\varnothing$.

Es obvio que $[0,1]\cup[2,3]$, o incluso $[0,1/2[\cup]1/2,1]$ son ejemplos de espacios disconexos. Notar que $[0,1/2[$ no es cerrado en $\mathbb{R}$, pero sí lo es en el espacio $[0,1/2[\,\cup\,]1/2,1]$ (su clausura en este espacio es la intersección con él de su clausura en $\mathbb{R}$, que es $[0,1/2]$, o sea, es $[0,1/2[$).

Es importante tener claro que los intervalos $[0,1/2[$ y $]1/2,1]$ están separados pese a que sólo falta un punto entre ellos. La falta de ese punto es suficiente para que ambas partes no se puedan "comunicar", en el sentido de que, por ejemplo, ninguna sucesión contenida en una de las piezas puede converger a un punto de la otra. Esto es suficiente para que ambas partes sean independientes topológicamente. Así, la función $f:[0,1/2[\,\cup\,]1/2,1]\longrightarrow\mathbb{R}$ dada por

$$f(x)=\begin{cases} 1 & \text{si } x\in[0,1/2[\,, \\ 2 & \text{si } x\in\,]1/2,1] \end{cases}$$

es continua, mientras que sería imposible definir una función continua sobre $[0,1]$ que sólo tomara los valores 1 y 2.

Si la desconexión de estos espacios es clara, no lo es tanto la conexión de espacios como $[0,1]$.

Ejercicio: Probar que el intervalo $[0,1]\subset\mathbb{Q}$ es disconexo.

Teorema 2.16 *Un subconjunto de $\overline{\mathbb{R}}$ es conexo si y sólo si es un intervalo.*

Demostración: Sea C un subespacio conexo de $\overline{\mathbb{R}}$. Sean a y b su ínfimo y su supremo, respectivamente. Vamos a probar que C es uno de los cuatro intervalos de extremos a y b. Para ello basta ver que si $a<x<b$ entonces $x\in C$. En caso contrario los conjuntos $C\cap[-\infty,x[$ y $C\cap\,]x,+\infty]$ son dos abiertos disjuntos no vacíos de C cuya unión es C.

Tomemos ahora un intervalo I y veamos que es conexo. Supongamos que existen abiertos disjuntos no vacíos U y V en I de modo que $I=U\cup V$. Tomemos $x\in U$ e $y\in V$. Podemos suponer que $x<y$.

Como I es un intervalo, $[x,y]\subset I$ y $U'=U\cap[x,y]$, $V'=V\cap[x,y]$ son abiertos disjuntos no vacíos en $[x,y]$ de modo que $[x,y]=U'\cup V'$.

Sea s el supremo de U'. Entonces $s\in\overline{U'}\cap[x,y]=U'$, luego en particular $s<y$. Claramente $]s,y]\subset V'$, luego $s\in\overline{V'}\cap[x,y]=V'$, contradicción. ■

Una consecuencia de esto es que un intervalo $[a,b[$ no es homeomorfo a uno de tipo $]c,d[$. En efecto, si eliminamos un punto de un intervalo $]c,d[$ nos queda un espacio disconexo, mientras que en $[a,b[$ podemos eliminar el punto a y obtenemos un conexo. (Si fueran homeomorfos, el espacio que resultara de eliminar la imagen de a en $]c,d[$ debería ser homeomorfo a $]a,b[$).

Ejercicio: Probar que dos intervalos (acotados o no acotados) son homeomorfos si y sólo si son del mismo tipo: abierto $]a,b[$, cerrado $[a,b]$ o semiabierto $[a,b[$.

Los resultados siguientes permiten probar con facilidad la conexión de muchos espacios. El primero refleja el hecho de que las aplicaciones continuas pueden pegar pero nunca cortar.

Teorema 2.17 *Las imágenes continuas de los espacios conexos son conexas.*

DEMOSTRACIÓN: Si $f : X \longrightarrow Y$ es una aplicación continua y suprayectiva pero Y no es conexo, entonces X tampoco puede serlo, pues si A es una abierto cerrado no vacío en Y y distinto de Y, entonces $f^{-1}[A]$ cumple lo mismo en X. ■

Teorema 2.18 *Sea $\{A_i\}_{i\in I}$ una familia de subespacios conexos de un espacio X tal que $\bigcap\limits_{i\in I} A_i \neq \varnothing$. Entonces $\bigcup\limits_{i\in I} A_i$ es conexo.*

DEMOSTRACIÓN: Supongamos que $\bigcup\limits_{i\in I} A_i = U \cup V$, donde U y V son abiertos disjuntos. Entonces para un i cualquiera se tendrá que $A_i = (U \cap A_i) \cup (V \cap A_i)$, pero $U \cap A_i$, $V \cap A_i$ son abiertos disjuntos en A_i, luego uno de ellos es vacío, y así $A_i \subset U$ o bien $A_i \subset V$.

Pero si $A_i \subset U$, entonces U contiene a $\bigcap\limits_{i\in I} A_i$, luego U corta a todos los A_i y por conexión los contiene a todos. Así $\bigcup\limits_{i\in I} A_i = U$, y $V = \varnothing$. Igualmente, si $A_i \subset V$ se deduce que U es vacío. ■

Ejemplo Las circunferencias son conexas.

Sea $f : [-1, 1] \longrightarrow \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1,\ y \geq 0\}$ la aplicación dada por $f(x) = \big(x, \sqrt{1 - x^2}\,\big)$.

Claramente f es continua y suprayectiva, lo que prueba que la semicircunferencia es conexa. Igualmente se prueba que la semicircunferencia opuesta es conexa, y como ambas se cortan en los puntos $(\pm 1, 0)$, su unión, es decir, la circunferencia, es conexa. ■

Teorema 2.19 *Si A es un subespacio conexo de un espacio X, entonces $\overline{A}$ es conexo.*

DEMOSTRACIÓN: Supongamos que $\overline{A} = U \cup V$, donde U y V son abiertos disjuntos en $\overline{A}$. Entonces $A = (U \cap A) \cup (V \cap A)$, y $U \cap A$, $V \cap A$ son abiertos disjuntos en A. Por conexión uno es vacío, luego $A \subset U$ o bien $A \subset V$. Digamos $A \subset U \subset \overline{A}$.

Pero U es cerrado en $\overline{A}$, luego $\overline{A} \subset \overline{U} = U$, es decir, $U = \overline{A}$ y $V = \varnothing$. Esto prueba que $\overline{A}$ es conexo. ■

Hemos dicho que un espacio disconexo es un espacio formado por varias "piezas" ahora podemos dar una definición rigurosa de lo que entendemos por una "pieza".

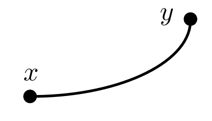
Definición 2.20 Sea X un espacio topológico y $x \in X$. Llamaremos *componente conexa* de x a la unión $C(x)$ de todos los subconjuntos conexos de X que contienen a x. Por el teorema 2.18, $C(x)$ es un conexo, el mayor subespacio conexo de X que contiene a x.

Es obvio que si $x, y \in X$, entonces $C(x)$ y $C(y)$ son iguales o disjuntas. En efecto, si tienen puntos en común, por el teorema 2.18 resulta que $C(x) \cup C(y)$ es un conexo, luego $C(x) \cup C(y) \subset C(x)$ y $C(x) \cup C(y) \subset C(y)$, con lo que $C(x) = C(x) \cup C(y) = C(y)$.

En resumen, todo espacio X está dividido en componentes conexas disjuntas. Las componentes conexas son cerradas por el teorema 2.19. En efecto, $\overline{C(x)}$ es un conexo que contiene a x, luego $\overline{C(x)} \subset C(x)$.

Sin embargo las componentes conexas no siempre son abiertas. Si un espacio tiene un número finito de componentes conexas, éstas serán abiertas y cerradas a la vez, evidentemente, pero si hay infinitas componentes ya no es necesario. Por ejemplo, ningún subconjunto de $\mathbb{Q}$ con más de un punto es conexo, porque no es un intervalo de $\mathbb{R}$, luego las componentes conexas de $\mathbb{Q}$ son los puntos, que no son abiertos.

A la hora de probar que un espacio es conexo, resulta útil el concepto de arco. Un *arco* en un espacio X es una aplicación continua $a : [0, 1] \longrightarrow X$. El espacio X es *arco-conexo* si para todo par de puntos $x, y \in X$ existe un arco $a : [0, 1] \longrightarrow X$ tal que $a(0) = x$, $a(1) = y$.



Como entonces x e y están en la imagen del arco a, que es un conexo, resulta que x e y están en la misma componente conexa de X, o sea, que X tiene una única componente conexa: Los espacios arco-conexos son conexos. El recíproco no es cierto, pero no vamos a dar un ejemplo.

Dados dos puntos $x, y \in \mathbb{R}^n$, el segmento que los une está formado por los puntos de la forma $y + \lambda(x - y)$, con $\lambda \in [0, 1]$. Esto se puede definir en cualquier $\mathbb{K}$-espacio vectorial. Si V es un espacio vectorial topológico y $x, y \in V$, entonces la aplicación $a : [0, 1] \longrightarrow V$ dada por $a(\lambda) = \lambda x + (1 - \lambda)y$ es un arco (el *segmento*) que une x con y.

Un subconjunto A de un $\mathbb{K}$-espacio vectorial V es *convexo* si para todos los puntos $x, y \in A$ y todo $\lambda \in [0, 1]$ se cumple $\lambda x + (1 - \lambda)y \in A$, es decir, si cuando A contiene a dos puntos, también contiene al segmento que los une.

Uniendo todo esto, resulta que en un espacio vectorial topológico, todo convexo es arco-conexo, luego conexo. En particular todo espacio vectorial topológico es conexo. En particular $\mathbb{K}^n$ es conexo.

Ejercicio: Probar que toda esfera de centro O en $\mathbb{R}^n$ es imagen continua de $\mathbb{R}^n \setminus \{0\}$. Probar que $\mathbb{R}^n \setminus \{0\}$ es conexo y deducir de aquí la conexión de la esfera.

Ejercicio: Probar que $\mathbb{R}^2$ no es homeomorfo a $\mathbb{R}$.

Los conjuntos convexos tienen una propiedad que en general no cumplen los conexos, y es que, claramente, la intersección de convexos es convexa.

Ejemplo Las bolas en los espacios normados son convexas.

En efecto, si x, $y \in B_\epsilon(z)$, entonces $\|x - z\| < \epsilon$, $\|y - z\| < \epsilon$, luego para todo $\lambda \in [0, 1]$ se cumple

$$\|\lambda x + (1 - \lambda)y - z\| = \|\lambda x + (1 - \lambda)y - \lambda z + (1 - \lambda)z\|$$

$$\leq \|\lambda(x - z)\| + \|(1 - \lambda)(y - z)\| = \lambda\|x - z\| + (1 - l)\|y - z\| < \lambda\epsilon + (1 - \lambda)\epsilon = \epsilon.$$

Por lo tanto $\lambda x + (1 - \lambda)y \in B_\epsilon(z)$.

(Cambiando las desigualdades estrictas por desigualdades no estrictas se prueba que las bolas cerradas son convexas.) ■

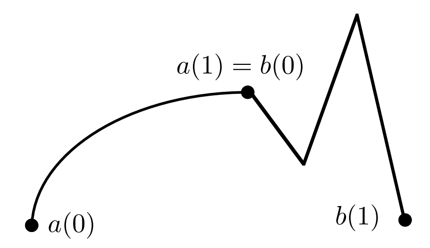
Por esto se dice que los espacios normados son localmente convexos. En general, un espacio vectorial topológico es *localmente convexo* si tiene una base formada por conjuntos convexos. Igualmente un espacio topológico es *localmente conexo* o *localmente arco-conexo* si tiene una base formada por abiertos conexos o arco-conexos, respectivamente. Así, los espacios localmente convexos son localmente arco-conexos y los espacios localmente arco-conexos son localmente conexos.

Un hecho importante es que si A es un abierto en un espacio localmente conexo X, entonces las componentes conexas de A son abiertas y cerradas. En efecto, si C es una componente conexa de A y $x \in C$, entonces existe un abierto (básico) U de X tal que U es conexo y $x \in U \subset A$. Como C es la componente conexa de x, ha de ser $U \subset C$, luego C es un entorno de x, o sea, C es entorno de todos sus puntos, luego es un abierto.

Ejercicio: Probar que todo abierto no vacío en $\mathbb{R}$ es la unión disjunta de una cantidad numerable de intervalos abiertos.

Veamos algunos hechos adicionales sobre arcos. Sean a y b dos arcos en un espacio X de modo que $a(1) = b(0)$.

Entonces la aplicación $b_1 : [1, 2] \longrightarrow X$ dada por $b_1(t) = b(t-1)$ es continua y cumple que $b_1(1) = b(0)$, $b_1(2) = b(1)$ (se trata de la composición con b del homeomorfismo $i : [1, 2] \longrightarrow [0, 1]$ dado por $i(t) = t - 1$).



Ahora, la unión $c = a \cup b_1 : [0, 2] \longrightarrow X$, esto es, la aplicación que restringida a $[0, 1]$ es a y restringida a $[1, 2]$ es b_1, es continua (porque restringida a los dos cerrados $[0, 1]$ y $[1, 2]$ lo es, y su imagen es la unión de las imágenes de a y b.

Finalmente, llamamos $a \cup b : [0, 1] \longrightarrow X$ a la función $(a \cup b)(t) = c(2t)$, es decir, la composición de c con el homeomorfismo $j : [0, 1] \longrightarrow [0, 2]$ definido

mediante $j(t) = 2t$. Claramente $a \cup b$ es un arco cuya imagen es la unión de las imágenes de a y b. En particular $(a \cup b)(0) = a(0)$ y $(a \cup b)(1) = b(1)$.

Esto significa que si podemos unir un punto x con un punto y a través de un arco a, y podemos unir un punto y con un punto z a través de un arco b, entonces podemos unir x con z mediante el arco $a \cup b$.

Por otra parte, si a es un arco en un espacio X, la aplicación $-a : [0, 1] \longrightarrow X$ dada por $(-a)(t) = a(1 - t)$ es un arco con la misma imagen pero de modo que $(-a)(0) = a(1)$ y $(-a)(1) = a(0)$. También es obvio que un arco constante une un punto consigo mismo.

De todo esto se sigue que la relación "x e y se pueden unir mediante un arco" es reflexiva, simétrica y transitiva en todo espacio X.

Teorema 2.21 *Sea X un espacio localmente arco-conexo. Entonces un abierto de X es conexo si y sólo si es arco-conexo.*

DEMOSTRACIÓN: Obviamente los abiertos arco-conexos son conexos. Supongamos que A es un abierto conexo no vacío. Sea $x \in A$. Sea U el conjunto de todos los puntos de A que pueden ser unidos con x mediante un arco contenido en A.

Veamos que U es abierto (en X o en A, es lo mismo). Sea $y \in U$. Entonces existe un arco a en A tal que $a(0) = x$, $a(1) = y$. Como X es localmente arco-conexo existe un abierto arco-conexo V tal que $y \in V \subset A$. Si $z \in V$, entonces hay un arco b en V (luego en A) que une y con z, luego $a \cup b$ es un arco en A que une x con z, luego $z \in U$.

Por lo tanto $V \subset U$ y así U es un entorno de y. Tenemos, pues, que U es entorno de todos sus puntos, luego es un abierto.

Ahora veamos que U es un cerrado en A, o lo que es lo mismo, que $A \setminus U$ es abierto. Como U no es vacío, por conexión tendrá que ser $U = A$, lo que significa que A es arco-conexo.

Si $y \in A \setminus U$, entonces y no puede ser unido a x mediante un arco. Existe un abierto arco-conexo V tal que $y \in V \subset A$, pero los puntos de V pueden unirse a y mediante un arco. Si alguno de estos puntos z pudiera unirse a x mediante un arco a, tendríamos un arco b que une a y con z y un arco a que une z con x, luego y se podría unir con x. Por lo tanto ningún punto de V puede unirse con x, es decir, $V \subset A \setminus U$, luego $A \setminus U$ es abierto. ■

Una poligonal es una unión de un número finito de segmentos. Una pequeña modificación del teorema anterior permite probar que en un espacio localmente convexo, un abierto es conexo si y sólo se es conexo por poligonales, es decir, todo par de puntos se puede unir por una poligonal.

Ahora vamos con las aplicaciones de la conexión. El resultado principal es el siguiente hecho obvio:

Teorema 2.22 (Teorema de los valores intermedios) *Si X es un espacio conexo, $f : X \longrightarrow \mathbb{R}$ es una aplicación continua, x, y son puntos de X y $f(x) < \alpha < f(y)$, entonces existe un punto $z \in X$ tal que $f(z) = \alpha$.*

DEMOSTRACIÓN: Se cumple que $f[X]$ es un conexo, luego un intervalo. Como $f(x)$ y $f(y)$ están en $f[X]$, a también ha de estar en $f[X]$. ■

A pesar de su simplicidad, las consecuencias de este teorema son importantes. Por ejemplo, no hay polinomios irreducibles de grado impar sobre $\mathbb{R}$, salvo los de grado 1:

Teorema 2.23 *Todo polinomio $p(x) \in \mathbb{R}[x]$ de grado impar tiene al menos una raíz en $\mathbb{R}$.*

DEMOSTRACIÓN: Como $p(x)$ tiene una raíz si y sólo si la tiene $-p(x)$, podemos suponer que su coeficiente director es positivo.

Entonces $\lim_{x\to+\infty} p(x) = +\infty$, mientras que $\lim_{x\to-\infty} p(x) = -\infty$. En particular existe un $u \in \mathbb{R}$ tal que $p(u) < 0$ y existe un $v \in \mathbb{R}$ tal que $p(v) > 0$. Por el teorema de los valores intermedios también existe un $a \in \mathbb{R}$ tal que $p(a) = 0$. ■

Por supuesto el teorema es falso para polinomios de grado par. Basta pensar en el caso $x^2 + 1$.

Si $a > 0$, el teorema de los valores intermedios aplicado al polinomio $x^n - a$ nos permite concluir la existencia de un $b > 0$ tal que $b^n = a$. Es claramente único, pues si $b^n = c^n$, entonces $(b/c)^n = 1$, de donde $b/c = \pm 1$, luego si ambos son positivos $b = c$.

Definición 2.24 Para cada natural $n > 0$ y cada número real $a > 0$ definimos la *raíz n-sima* de a como el único número $b > 0$ tal que $b^n = a$. Lo representaremos $b = \sqrt[n]{a}$

Unas comprobaciones rutinarias muestran que si m, n son números enteros $n > 0$ y $a > 0$ entonces el número $a^{m/n} = \left(\sqrt[n]{a}\right)^m$ depende sólo de la fracción m/n, con lo que tenemos definida la exponencial a^r para todo número real positivo a y todo número racional r y extiende a la exponencial entera. También se comprueba que $a^{r+s} = a^r a^s$, $(a^r)^s = a^{rs}$.

***Afinidades directas e inversas** Las isometrías de un espacio afín euclídeo E se clasifican en movimientos y simetrías según que el determinante de la aplicación lineal asociada sea igual a 1 o a -1. La topología proporciona una interpretación geométrica de esta distinción puramente algebraica. En efecto, es conocido que todo movimiento (en dimensión mayor que 1) se descompone en composición de giros, y un giro es una aplicación f que en un sistema de referencia afín adecuado tiene la expresión:

$$\begin{aligned} x_1' &= x_1 \cos\alpha - x_2 \operatorname{sen}\alpha, \\ x_2' &= x_1 \operatorname{sen}\alpha + x_2 \cos\alpha \\ x_3' &= x_3, \\ \cdots & \quad \cdots \\ x_n' &= x_n \end{aligned}$$

Vamos a admitir la continuidad de las funciones trigonométricas. La demostraremos en el capítulo III.

Consideremos ahora la aplicación $f : [0,1] \times E \longrightarrow E$ tal que si el punto x tiene coordenadas $(x_1, \ldots, x_n)$ en el mismo sistema de referencia, entonces $f_t(x) = f(t,x)$ es el punto de coordenadas $(x'_1, \ldots, x'_n)$ dadas por

$$\begin{array}{rcl} x'_1 & = & x_1 \cos t\alpha - x_2 \operatorname{sen} t\alpha, \\ x'_2 & = & x_1 \operatorname{sen} t\alpha + x_2 \cos t\alpha \\ x'_3 & = & x_3, \\ \cdots & & \cdots \\ x'_n & = & x_n \end{array}$$

Claramente entonces, para cada $t \in [0,1]$, la aplicación f_t es un giro de ángulo $t\alpha$, de modo que f_0 es la identidad y $f_1 = f$. Además la aplicación f es continua. Veamos que podemos obtener una aplicación similar para movimientos arbitrarios:

Teorema 2.25 *Si f es un movimiento en un espacio afín euclídeo E existe una aplicación continua $f : [0,1] \times E \longrightarrow E$ tal que para todo $t \in [0,1]$, la aplicación $f_t(x) = f(t,x)$ es un movimiento, $f_0 = 1$ y $f_1 = f$.*

Demostración: Lo probamos por inducción sobre el número de giros en que se descompone f. Ya lo tenemos probado cuando f es un giro. Basta probar que si f cumple el teorema, g es un giro y $h = fg$ entonces h cumple el teorema. Tenemos, pues, las aplicaciones f_t y g_t. Sea $h : [0,1] \times E \longrightarrow E$ dada por

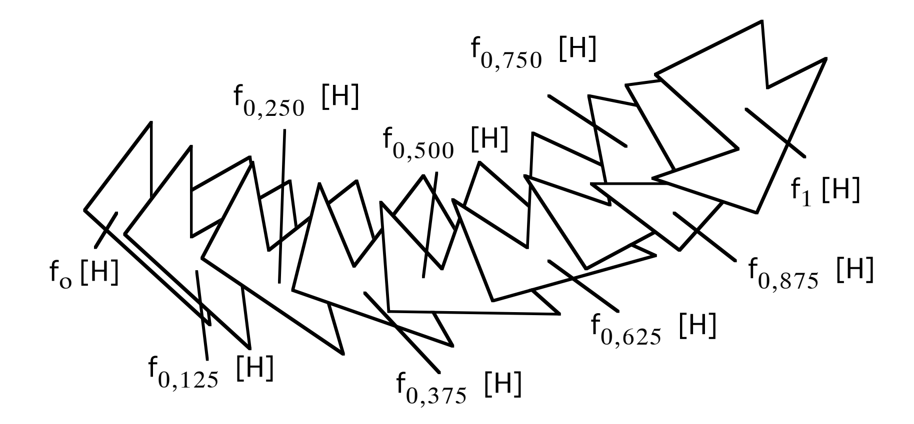
$$h(t,x) = \begin{cases} f(2t,x) & \text{si } 0 \leq t \leq 1/2, \\ g\big(2t-1, f(x)\big) & \text{si } 1/2 < t \leq 1, \end{cases}$$

La aplicación h es claramente continua en el cerrado $[0,1/2] \times E$. Basta observar que su restricción al cerrado $[1/2,1] \times E$ viene dada por $g\big(2t-1, f(x)\big)$, pues entonces también será continua en este cerrado y por consiguiente en todo su dominio. Ahora bien, los únicos puntos donde esta igualdad no es cierta por definición son los de la forma $(1/2, x)$, pero

$$h(1/2,x) = f(1,x) = f(x) = g\big(0, f(x)\big) = g\big(2 \cdot (1/2) - 1, f(x)\big).$$

El resto del teorema es obvio: para cada t la aplicación h_t es de la forma $f_{t'}$ o $g_{t'}$, luego es un movimiento, etc. ■

Este teorema se interpreta como que los movimientos pueden efectuarse de forma continua en el tiempo. El punto $f_t(x)$ se interpreta como la posición x en el instante t, de modo que para $t = 0$ tenemos la posición inicial x y para $t = 1$ tenemos la posición final $f(x)$. El arco $f(t,x)$, para un x fijo, es la trayectoria que sigue x. El hecho de que cada f_t sea un movimiento se interpreta como que en cada instante t las distancias entre los puntos son las mismas que las originales.



Ahora probamos que esta propiedad distingue a los movimientos de las semejanzas.

Teorema 2.26 *Si E es un espacio afín euclídeo y $f : [0,1] \times E \longrightarrow E$ es una aplicación continua tal que para todo $t \in [0,1]$, la aplicación $f_t(x) = f(t,x)$ es una isometría y $f_0 = 1$, entonces todas las aplicaciones f_t son movimientos.*

DEMOSTRACIÓN: Fijado un sistema de referencia afín en E, la aplicación $h : E \longrightarrow \mathbb{R}^n$ que a cada punto le asigna sus coordenadas es un homeomorfismo. La aplicación $g : [0,1] \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ dada por $g(t,X) = h(f(t,h^{-1}(X))$ es continua, y es de la forma $g(t,X) = P_t + XA_t$, donde $P_t \in \mathbb{R}^n$ y A_t es una matriz $n \times n$ de determinante ± 1.

La aplicación $g'(t,X) = g(t,X) - g(t,0) = XA_t$ también es continua. Si e_i es el vector i-ésimo de la base canónica, la aplicación $g_{ij}(t) = g'(t,e_i)e_j$ es continua, y $g_{ij}(t)$ es simplemente el coeficiente (i,j) de la matriz A_t. El determinante de una matriz depende polinómicamente de sus coeficientes, por lo que la función $\det f_t = \det A_t$ es continua.

Ahora bien, si alguna f_t fuera una simetría, $\det f_0 = \det I = 1$, mientras que $\det f_t = \det A_t = -1$. Tendrámos, pues, una aplicación continua y suprayectiva del espacio conexo $[0,1]$ en el espacio disconexo $\{-1,1\}$, lo cual es imposible.

■

Los mismos argumentos sirven para interpretar otros grupos de afinidades. Por ejemplo:

Teorema 2.27 *Si f es una biyección afín de determinante positivo en un espacio afín E existe una aplicación continua $f : [0,1] \times E \longrightarrow E$ tal que para todo $t \in [0,1]$, la aplicación $f_t(x) = f(t,x)$ es una biyección afín, $f_0 = 1$ y $f_1 = f$.*

DEMOSTRACIÓN: Probaremos primero el teorema para automorfismos de determinante positivo del espacio vectorial asociado $\vec{E}$. Cada uno de estos automorfismos se puede expresar como composición de una homotecia lineal de determinante positivo y un automorfismo de determinante 1. A su vez, estos automorfismos se descomponen en producto de transvecciones, es decir,

de aplicaciones de la forma $f(x) = x + u(x)h$, donde $u : \vec{E} \longrightarrow \mathbb{R}$ es una aplicación lineal y h es un vector del núcleo de u.

Para una transvección definimos

$$f(t,x) = x + tu(x)h,$$

que es una transvección para todo t, es continua como aplicación $[0,1]\times\vec{E} \longrightarrow \vec{E}$, $f_0 = 1$ y $f_1 = f$.

Una homotecia lineal es de la forma $f(x) = rx$. Definimos la aplicación $f(t,x) = \big(1 + t(r-1)\big)x$. Como $1 + t(r-1) > 0$ siempre que $r > 0$ y $0 \leq t \leq 1$, tenemos que f_t es una homotecia lineal para todo t. Claramente se cumplen también las otras propiedades.

Aplicando la misma técnica de composición que en el caso de los movimientos llegamos a que todo automorfismo de $\vec{E}$ de determinante positivo cumple el teorema.

Una biyección afín en E de determinante positivo es de la forma

$$f(P) = O + \vec{v} + \vec{f}(\overrightarrow{OP}),$$

donde O es un punto arbitrario de E y $\vec{f}$ es un automorfismo de $\vec{E}$ de determinante positivo. Definimos $f(t,P) = O + t\vec{v} + \vec{f}_t(\overrightarrow{OP})$. Es fácil ver que esta aplicación cumple lo pedido. ■

El teorema es falso para biyecciones afines de determinante negativo, pues con el mismo argumento que hemos empleado para las simetrías se llega a que la aplicación $\det f_t$ es una aplicación continua en el espacio conexo $[0,1]$ tal que $\det f_0 = 1$, $\det f_1 < 0$ pero que nunca toma el valor 0, lo cual es imposible. Si un endomorfismo de $\vec{E}$ tiene determinante 0 su imagen tiene dimensión menor que $\vec{E}$. Por consiguiente, cualquier aplicación que transforme continuamente un conjunto en su imagen por una biyección afín de E de determinante negativo, en un momento dado "aplanará" el espacio en una variedad afín de dimensión menor. ■

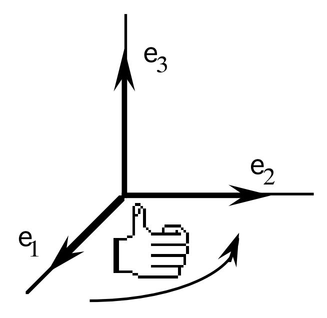
Orientación Los resultados que acabamos de ver nos llevan a introducir un concepto de orientación en un espacio vectorial (aunque casi todo el razonamiento que sigue es independiente de lo anterior).

Definición 2.28 Diremos que dos bases ordenadas de un espacio vectorial real de dimensión finita V tienen la misma *orientación* si el determinante de la matriz de cambio de base es positivo. Alternativamente, si existe un isomorfismo de determinante positivo que transforma una en otra.

Es inmediato comprobar que las bases de V se dividen en dos clases de equivalencia, a las que llamaremos *orientaciones* de V. Si una base tiene una determinada orientación, al cambiar de signo uno cualquiera de sus vectores pasamos a una base con la orientación opuesta.

Un espacio vectorial *orientado* es un espacio vectorial real de dimensión finita en el que hemos seleccionado una orientación, a la que llamaremos *positiva,* mientras que a la orientación opuesta la llamaremos *negativa.* Consideraremos a $\mathbb{R}^n$ como espacio orientado tomando como orientación positiva a la que contiene a la base canónica.

Conviene notar que en otros espacios vectoriales, por ejemplo en los subespacios de $\mathbb{R}^n$, no hay ninguna base privilegiada que nos permita definir una orientación, luego la elección de una u otra como positiva es arbitraria. Lo mismo sucede con el espacio vectorial asociado a un espacio afín real, o con el espacio intuitivo, donde no tenemos bases canónicas. Para determinar una orientación en las representaciones gráficas hemos de recurrir a criterios no geométricos. Por ejemplo, si representamos la recta horizontalmente, se suele considerar como base positiva a la formada por el único vector unitario que apunta hacia la derecha. La distinción izquierda-derecha no tiene más fundamento que la anatomía humana.



Para adoptar criterios similares de orientación en el plano y el espacio debemos notar primero que, según los resultados de la sección anterior, dos bases ortonormales tienen la misma orientación si y sólo si podemos transformar una en otra mediante un movimiento continuo en el tiempo. Así, diremos que una base del plano es positiva si cuando el dedo índice derecho apunta en la dirección (y sentido) del primer vector (con la palma hacia abajo) entonces el pulgar (puesto en ángulo recto) apunta en la dirección del segundo vector. Si dos bases cumplen esto, el movimiento que transporta nuestra mano de una a la otra justifica que tienen la misma orientación, y viceversa. Similarmente, consideraremos positivas a las bases ortonormales del espacio tales que podemos disponer la mano derecha con el dedo medio apuntando en la dirección del primer vector, el pulgar en la del segundo y el índice en la del tercero. Una forma equivalente y más cómoda de esta regla es la siguiente: Una base es positiva si cuando el índice derecho arqueado marca el sentido de giro que lleva del primer vector al segundo por el ángulo más corto, entonces el pulgar apunta en la dirección del tercer vector. Una ligera modificación de estas reglas las hace válidas para bases cualesquiera, no necesariamente ortonormales.

Si un vector v en un plano orientado tiene coordenadas (a, b) respecto a una base ortonormal positiva, entonces es claro que el vector w de coordenadas $(-b, a)$ es ortogonal al primero, con el mismo módulo y la base (v, w) es positiva. Vamos a obtener un resultado análogo en tres dimensiones.

Sean v y w dos vectores en un espacio tridimensional orientado V. Sean (a_1, a_2, a_3), (b_1, b_2, b_3) sus coordenadas en una base (e_1, e_2, e_3) ortonormal y positiva. Definimos su *producto vectorial* $v \wedge w$ como el vector cuyas coordenadas en dicha base son

$$\left(\begin{vmatrix} a_1 & a_3 \\ b_2 & b_3 \end{vmatrix}, \begin{vmatrix} a_3 & a_1 \\ b_3 & b_1 \end{vmatrix}, \begin{vmatrix} a_1 & a_2 \\ b_1 & b_1 \end{vmatrix} \right).$$

Como regla mnemotécnica y de cálculo podemos usar:

$$v \wedge w = \begin{vmatrix} e_1 & e_2 & e_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix}.$$

Hemos de probar que esta definición no depende de la elección de la base, siempre que sea positiva. Aunque con rigor el determinante anterior no tiene sentido (pues tiene vectores en su primera fila), sí hay una relación sencilla entre el producto vectorial y los determinantes que es útil para deducir sus propiedades. Si x es un vector de coordenadas (x_1, x_2, x_3), entonces

$$(x, u, v) = x(v \wedge w) = \begin{vmatrix} x_1 & x_2 & x_3 \\ a_1 & a_2 & a_3 \\ b_1 & b_2 & b_3 \end{vmatrix},$$

donde este determinante sí tiene sentido. El escalar (x, u, v) se llama *producto mixto* de los vectores x, u, v, y es claro que no depende de la base ortonormal positivamente orientada que se escoja para calcularlo. Teniendo esto en cuenta es fácil probar hechos como que $v \wedge w = -w \wedge v$. En efecto, para todo x se cumple evidentemente $x(v \wedge w) = -x(w \wedge v)$, y esto sólo es posible si se da la relación indicada. Del mismo modo se prueban las relaciones

$$v \wedge (w + x) = v \wedge w + v \wedge x, \quad \alpha v \wedge w = (\alpha v) \wedge w = v \wedge (\alpha w), \quad \alpha \in \mathbb{R}.$$

Además $v \wedge w = 0$ si y sólo si v y w son linealmente dependientes. En efecto, si son dependientes $x(v \wedge w) = 0$ para todo x, luego $v \wedge w = 0$. Si son independientes entonces existe un x independiente de ambos, de modo que $x(v \wedge w) \neq 0$, luego $v \wedge w \neq 0$.

Si v y w son linealmente independientes, entonces $v \wedge w$ es un vector ortogonal a ambos. En efecto, $v(v \wedge w) = w(v \wedge w) = 0$. Más aún, en general se cumple

$$\|v \wedge w\| = \|v\| \, \|w\| \operatorname{sen} \widehat{vw}.$$

Para probarlo basta comprobar la identidad

$$\|v \wedge w\|^2 + (vw)^2 = \|v\|^2 \, \|w\|^2,$$

que junto con $vw = \|v\| \, \|w\| \cos \widehat{vw}$ nos lleva a la relación indicada.

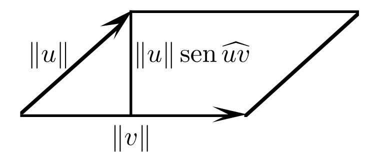
Por último observamos que si v y w son linealmente independientes entonces la base $(v, w, v \wedge w)$ es positiva, pues el determinante de la matriz de cambio de base respecto a e_1, e_2, e_3 es

$$\begin{vmatrix} a_1 & a_3 \\ b_2 & b_3 \end{vmatrix}^2 + \begin{vmatrix} a_3 & a_1 \\ b_3 & b_1 \end{vmatrix}^2 + \begin{vmatrix} a_1 & a_2 \\ b_1 & b_1 \end{vmatrix}^2 > 0.$$

Estas propiedades muestran que $v \wedge w$ es independiente de la base respecto a la cual lo calculamos. Sea cual sea esta base, el producto vectorial resulta ser

el vector nulo si v y w son dependientes o bien el vector perpendicular a ambos cuyo módulo es el que hemos calculado y cuyo sentido es el necesario para que la base $(v, w, v \wedge w)$ sea positiva.

Para terminar daremos una interpretación intuitiva del módulo del producto vectorial de dos vectores u y v. Se trata claramente del área del paralelogramo que los tiene por lados:



2.3 Espacios completos

La última propiedad que vamos a estudiar no es topológica, sino métrica. La completitud garantiza la convergencia de ciertas sucesiones sin necesidad de conocer su límite de antemano. Ya hemos encontrado algunos casos, como el de las sucesiones monótonas en $\overline{\mathbb{R}}$, o las monótonas y acotadas en $\mathbb{R}$. Los criterios de este son muy útiles porque permiten definir nuevas funciones y constantes como límites de sucesiones. Comenzamos introduciendo una familia de sucesiones en un espacio métrico que incluye a todas las convergentes:

Definición 2.29 Sea M un espacio métrico. Una sucesión $\{a_n\}_{n=0}^{\infty}$ en M es una sucesión *de Cauchy* si para todo $\epsilon > 0$ existe un natural n_0 de modo que si m, $n \geq n_0$, entonces $d(a_m, a_n) < \epsilon$.

O sea, una sucesión es de Cauchy si sus términos están finalmente tan próximos entre sí como se desee. Esto le ocurre a toda sucesión convergente, pues si $\{a_n\}_{n=0}^{\infty}$ converge a L, entonces dado $\epsilon > 0$ existe un natural n_0 de modo que para todo $n \geq n_0$ se cumple $d(a_n, L) < \epsilon/2$, luego si m, $n \geq n_0$ se cumple

$$d(a_m, a_n) \leq d(a_m, L) + d(L, a_n) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Así pues, toda sucesión convergente es de Cauchy, pero el recíproco no es cierto. Pensemos en la sucesión $\{1/n\}$ en el espacio $]0, 1]$. Como en $\mathbb{R}$ es convergente, es de Cauchy. Obviamente sigue siendo de Cauchy como sucesión en $]0, 1]$, pero ya no es convergente.

Por supuesto que el problema no está en la sucesión, sino en el espacio, al que en cierto sentido "le falta un punto". Los espacios a los que no les faltan puntos en este sentido se llaman completos:

Un espacio métrico M es *completo* si toda sucesión de Cauchy en M es convergente.

La completitud de $\mathbb{R}$ es consecuencia de las siguientes propiedades obvias de las sucesiones de Cauchy:

Teorema 2.30 *Sea M un espacio métrico y $\{a_n\}_{n=0}^{\infty}$ una sucesión de Cauchy.*

a) Si $\{a_n\}_{n=0}^{\infty}$ *tiene un punto adherente* L*, entonces converge a* L*.*

b) La sucesión $\{a_n\}_{n=0}^{\infty}$ *está acotada.*

DEMOSTRACIÓN: a) Dado $\epsilon > 0$, existe un número natural a partir del cual $d(a_m, a_n) < \epsilon/2$, y existe también un natural n_0, que podemos tomar mayor que el anterior, tal que $d(a_{n_0}, L) < \epsilon/2$. Por lo tanto si $n \geq n_0$ se cumple que

$$d(a_n, L) \leq d(a_n, a_{n_0}) + d(a_{n_0}, L) < \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon.$$

Por consiguiente la sucesión converge a L.

b) Existe un n_0 tal que si $n \geq n_0$, entonces $d(a_n, a_{n_0}) < 1$, esto significa que $\{a_n \mid n \geq n_0\} \subset B_1(a_{n_0})$, luego es un conjunto acotado, y la sucesión completa es la unión de este conjunto con el conjunto de los n_0 primeros términos, que es finito, luego también está acotado. Por lo tanto la sucesión está acotada. ■

Teorema 2.31 *Todo espacio normado de dimensión finita es completo.*

DEMOSTRACIÓN: Una sucesión de Cauchy está acotada, luego está contenida en una bola cerrada, que es un conjunto compacto, luego tiene un punto adherente, luego converge. ■

En la sencilla prueba de este teorema se ve una relación entre la compacidad y la completitud. Con más detalle, la situación es la siguiente:

Teorema 2.32 *Se cumple:*

a) Todo espacio métrico compacto es completo.

b) Todo cerrado en un espacio métrico completo es completo.

c) Todo subespacio completo de un espacio métrico es cerrado.

DEMOSTRACIÓN: a) Toda sucesión tiene una subsucesión convergente, luego si es de Cauchy es convergente.

b) Si M es un espacio métrico completo y C es un cerrado en M, entonces toda sucesión de Cauchy en C converge en M, y como C es cerrado su límite estará en C, luego la sucesión converge en C.

c) Si C es un subespacio completo de un espacio métrico M, dado un punto x en la clausura de C, existe una sucesión en C que converge a x, luego la sucesión es de Cauchy, luego converge en C, luego x está en C, luego C es cerrado. ■

Por supuesto no todo espacio completo es compacto. A veces es útil conocer lo que separa a un espacio completo de la compacidad:

Definición 2.33 Un espacio métrico M es *precompacto* si para cada $\epsilon > 0$ existen puntos $x_1, \ldots, x_n$ en M tales que $M = B_\epsilon(x_1) \cup \cdots \cup B_\epsilon(x_n)$.

Obviamente todo espacio compacto es precompacto (basta extraer un subcubrimiento finito del cubrimiento formado por todas las bolas de radio ϵ).

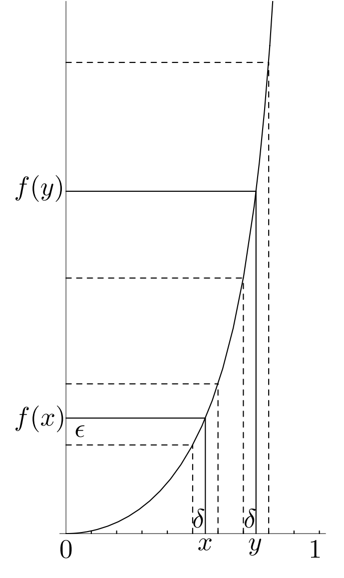
El teorema 2.7 contiene la demostración de que un espacio precompacto y completo es compacto. En efecto partiendo de la hipótesis sobre subsucesiones convergentes, en primer lugar se prueba que el espacio es precompacto. Con ayuda de la precompacidad se construye una sucesión de bolas $B_{1/(n+1)}(x_n)$ en las que podemos exigir que cada una de ellas corte a la anterior. Esto garantiza que la sucesión de los centros es de Cauchy, con lo que podemos garantizar su convergencia por la completitud y probamos la compacidad del espacio. Así pues:

Teorema 2.34 *Un espacio métrico es compacto si y sólo si es precompacto y completo.*

Es muy importante tener claro que la completitud es una propiedad métrica y no topológica. Por ejemplo, los espacios $\mathbb{R}$ y $]-1,1[$ son homeomorfos, pero uno es completo y el otro no. En particular una imagen continua de un espacio completo no tiene por qué ser completo.

Si analizamos lo que falla, vemos que en $]-1,1[$ hay sucesiones de Cauchy no convergentes, por ejemplo las que convergen a 1 en $\mathbb{R}$, pero cuando las transformamos por el homeomorfismo entre $]-1,1[$ y $\mathbb{R}$ se convierten en sucesiones que tienden a $+\infty$, que ya no son de Cauchy, luego no violan la completitud de $\mathbb{R}$. El problema es que mientras una aplicación continua transforma sucesiones convergentes en sucesiones convergentes, no transforma necesariamente sucesiones de Cauchy en sucesiones de Cauchy. Y a su vez esto se debe a que si se estira infinitamente una sucesión de Cauchy, ésta deja de serlo. Esto nos lleva a conjeturar que la completitud se conservará por aplicaciones continuas que no produzcan estiramientos infinitos. Vamos a definir este tipo de aplicaciones.

Definición 2.35 Una aplicación $f : M \longrightarrow N$ entre espacios métricos es *uniformemente continua* si para todo $\epsilon > 0$ existe un $\delta > 0$ de modo que si $x, y \in M$ cumplen $d(x,y) < \delta$, entonces $d\big(f(x), f(y)\big) < \epsilon$.



Desde un punto de vista lógico, la diferencia entre aplicación continua y uniformemente continua es sutil. Conviene confrontarlas. Recordemos que f es continua en un punto x si para todo $\epsilon > 0$ existe un $\delta > 0$ tal que si $y \in M$ cumple $d(x,y) < \delta$, entonces $d\big(f(x), f(y)\big) < \epsilon$.

La diferencia es, pues, que cuando f es uniformemente continua el mismo δ verifica la definición de continuidad para un ϵ dado simultáneamente en todos los puntos x. Por ejemplo, el homeomorfismo f entre $]-1,1[$ y $\mathbb{R}$ estira más los puntos cuanto más próximos están de ± 1. Si tomamos

un punto x y queremos garantizar que $f(z)$ diste de $f(x)$ menos que ϵ, tendremos que exigir que z diste de x menos que un cierto δ, pero si consideramos los puntos que distan menos que δ de un punto y más cercano a 1, vemos que muchos de ellos se transforman en puntos que distan de $f(y)$ mucho más que ϵ, y que si queremos que no sobrepasen esta cota hemos de tomar un δ mucho menor.

Teorema 2.36 *Sea $f : K \longrightarrow M$ una aplicación continua entre espacios métricos y supongamos que K es compacto. Entonces f es uniformemente continua.*

DEMOSTRACIÓN: Sea $\epsilon > 0$. Para cada punto $x \in K$, como f es continua en x, existe un $\delta(x) > 0$ tal que si $d(y, x) < \delta(x)$, entonces $d\big(f(y), f(x)\big) < \epsilon/2$. Cubramos el espacio K por todas las bolas abiertas de centro cada punto x y de radio $\delta(x)/2$ y tomemos un subcubrimiento finito. Digamos que las bolas que forman este subcubrimiento tienen centros en los puntos $x_1, \ldots, x_n$. Sea $\delta > 0$ el mínimo del conjunto $\{\delta(x_1)/2, \ldots, \delta(x_n)/2\}$.

Si x, y son dos puntos cualesquiera de K tales que $d(x, y) < \delta$, entonces x estará en una de las bolas $B_{\delta(x_i)/2}(x_i)$. Entonces

$$d(y, x_i) \leq d(y, x) + d(x, x_i) < \frac{\delta(x_i)}{2} + \frac{\delta(x_i)}{2} = \delta(x_i).$$

Por lo tanto x, $y \in B_{\delta(x_i)}(x_i)$, de donde

$$d\big(f(x), f(x_i)\big) < \frac{\epsilon}{2}, \quad d\big(f(y), f(x_i)\big) < \frac{\epsilon}{2}.$$

Consecuentemente, $d\big(f(x), f(y)\big) < \epsilon$. ■

La imagen de una sucesión de Cauchy por una aplicación uniformemente continua es una sucesión de Cauchy. En efecto, si $\{a_n\}_{n=0}^{\infty}$ es de Cauchy y f es uniformemente continua, dado $\epsilon > 0$ existe un $\delta > 0$ de manera que si $d(x, y) < \delta$, entonces $d\big(f(x), f(y)\big) < \epsilon$. Existe un natural no tal que si n, $m \geq n_0$ entonces $d(a_m, a_n) < \delta$, luego $d\big(f(a_m), f(a_n)\big) < \epsilon$. Esto prueba que la sucesión $\{f(a_n)\}_{n=0}^{\infty}$ es de Cauchy.

Como consecuencia, si dos espacios métricos X e Y son *uniformemente homeomorfos,* esto es, si existe una biyección uniformemente continua con inversa uniformemente continua entre ellos, uno es completo si y sólo si lo es el otro.

Es importante destacar que una imagen uniformemente continua de un espacio completo no tiene por qué ser completa.

Veamos ahora algunas propiedades de las aplicaciones lineales entre espacios normados.

Teorema 2.37 *Sea $f : E \longrightarrow F$ una aplicación lineal entre espacios normados. Las siguientes condiciones son equivalentes:*

a) f es continua en E.

b) f es continua en 0.

c) f está acotada en $\overline{B}_1(0)$.

d) Existe un $M \geq 0$ tal que para todo $x \in E$ se cumple $\|f(x)\| \leq M\|x\|$.

e) f es uniformemente continua en E.

Demostración: a) → b) es obvio.

b) → c), pues existe un $\delta > 0$ tal que $\|x-0\| \leq \delta$, entonces $\|f(x)-f(0)\| \leq 1$.

Por lo tanto si $\|x\| \leq 1$, se cumple $\|\delta\, x\| \leq \delta$, $\|f(\delta\, x)\| \leq 1$, $\|f(x)\| \leq 1/\delta$, o sea, que $1/\delta$ es una cota de f en $\overline{B}_1(0)$.

c) → d), pues si M es una cota de f en $\overline{B}_1(0)$, dado cualquier $x \neq 0$ se cumple que $x/\|x\| \in \overline{B}_1(0)$, luego $\|f(x/\|x\|)\| \leq M$, de donde $\|f(x)\| \leq M\|x\|$, y esto también es cierto si $x = 0$.

d) → e), pues para todos los x, y en E:

$$\|f(x) - f(y)\| = \|f(x-y)\| \leq M\|x-y\|,$$

luego dado $\epsilon > 0$, si $\|x-y\| < \epsilon/M$, se cumple $\|f(x) - f(y)\| < \epsilon$.

e) → a) es evidente. ■

En particular, todo isomorfismo entre dos $\mathbb{K}$-espacios vectoriales de dimensión finita es un homeomorfismo uniforme para cualquier par de normas.

Definición 2.38 Un *espacio de Banach* es un espacio normado completo.

Hemos visto que todo espacio normado de dimensión finita es un espacio de Banach.

2.4 Espacios de Hilbert

En el capítulo I introdujimos los espacios prehilbertianos, que son los $\mathbb{K}$-espacios vectoriales dotados de un producto escalar.

Definición 2.39 Un *espacio de Hilbert* es un espacio prehilbertiano H que sea completo con la métrica inducida por el producto escalar.

Vamos a probar algunos hechos de interés sobre espacios de Hilbert. Los primeros son válidos en general sobre espacios prehilbertianos:

Teorema 2.40 *Si H es un espacio prehilbertiano, entonces el producto escalar en H es una función continua.*

Demostración: Para todos los x, x', y, $y' \in H$ se cumple

$$\begin{aligned} |x\cdot y - x'\cdot y'| &\leq |(x-x')\cdot y| + |x'\cdot(y-y')| \leq \|x-x'\|\,\|y\| + \|x'\|\,\|y-y'\| \\ &\leq \|x-x'\|\,\|y\| + \|x'-x\|\,\|y-y'\| + \|x\|\,\|y-y'\|. \end{aligned}$$

Así, dado un par $(x, y) \in H \times H$ y un $\epsilon > 0$, todo par $(x', y') \in H \times H$ que cumpla $\|x'-x\|$, $\|y'-y\| < \epsilon/3M$, donde $M > \|x\|, \|y\|$, cumple también que

$$|x\cdot y - x'\cdot y'| < \frac{\epsilon}{3M}\,M + \frac{\epsilon^2}{9M^2} + \frac{\epsilon}{3M}\,M < \epsilon.$$

■

Definición 2.41 Si H es un espacio prehilbertiano, diremos que $x, y \in H$ son *ortogonales*, y lo representaremos por $x \perp y$, si $x \cdot y = 0$.

Es conocido que la ortogonalidad en $\mathbb{R}^n$ coincide con el concepto geométrico de perpendicularidad. Es fácil generalizar el teorema de Pitágoras:

Si $x \perp y$, *entonces* $\|x+y\|^2 = \|x\|^2 + \|y\|^2$.

Así mismo, las propiedades del producto escalar dan inmediatamente la fórmula conocida como *identidad del paralelogramo:*

$$\|x+y\|^2 + \|x-y\|^2 = 2\|x\|^2 + 2\|y\|^2.$$

Para cada $A \subset H$, definimos

$$A^\perp = \{x \in H \mid x \perp a \text{ para todo } a \in A\}.$$

Es claro que $A^\perp$ es un subespacio vectorial de H. Más aún, puesto que $\{a\}^\perp$ es la antiimagen de 0 por la aplicación continua $x \mapsto a \cdot x$, se cumple que $\{a\}^\perp$ es cerrado, y como

$$A^\perp = \bigcap_{a\in A} \{a\}^\perp,$$

vemos que $A^\perp$ es un subespacio cerrado de H.

Si V es un subespacio vectorial de H es claro que $V \cap V^\perp = 0$. Vamos a probar que si V es cerrado entonces $H = V \oplus V^\perp$. Para ello necesitamos un resultado previo:

Teorema 2.42 *Sea M un subconjunto no vacío, cerrado y convexo de un espacio de Hilbert H. Entonces M contiene un único elemento de norma mínima.*

Demostración: Sea δ el ínfimo de las normas de los elementos de M. Aplicando la identidad del paralelogramo a $\frac{1}{2}x$, $\frac{1}{2}y$ tenemos

$$\frac{1}{4}\|x-y\|^2 = \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \left\|\frac{x+y}{2}\right\|^2.$$

Si $x, y \in M$, por convexidad $(x+y)/2 \in M$, luego

$$\|x-y\|^2 \leq 2\|x\|^2 + 2\|y\|^2 - 4\delta^2. \tag{2.1}$$

Si $\|x\| = \|y\| = \delta$ esto implica $x = y$, lo que nos da la unicidad. Es fácil construir una sucesión $\{x_n\}_{n=0}^\infty \subset M$ tal que $\lim\limits_n \|x_n\| = \delta$. Aplicando 2.1 a x_m y x_n concluimos fácilmente que la sucesión es de Cauchy, luego converge a un punto x que, por continuidad de la norma, cumplirá $\|x\| = \delta$. Además, como M es cerrado, ha de ser $x \in M$ y claramente su norma es la mínima en M. ■

Teorema 2.43 *Sea V un subespacio cerrado de un espacio de Hilbert H. Entonces*

a) *Todo $x \in H$ se descompone de forma única como $x = Px + Qx$, donde $Px \in V$, $Qx \in V^\perp$.*

b) *Px y Qx son los puntos de V y $V^\perp$ más próximos a x.*

c) *Las aplicaciones $P : H \longrightarrow V$ y $Q : H \longrightarrow V^\perp$ son lineales y continuas.*

d) $\|x\|^2 = \|Px\|^2 + \|Qx\|^2$.

Demostración: El conjunto $x + V$ es cerrado y convexo, luego podemos definir Qx como el elemento de norma mínima en $x + V$. Definimos $Px = x - Qx$. Obviamente $Px \in V$. Veamos que $Qx \in V^\perp$. Para ello probaremos que $(Qx) \cdot y = 0$ para todo $y \in V$. No perdemos generalidad si suponemos $\|y\| = 1$. Por definición de Qx tenemos que

$$(Qx) \cdot (Qx) = \|Qx\|^2 \leq \|Qx - \alpha y\|^2 = (Qx - \alpha y) \cdot (Qx - \alpha y),$$

para todo $\alpha \in \mathbb{K}$. Simplificando queda

$$0 \leq -\alpha(y \cdot Qx) - \bar{\alpha}(Qx \cdot y) + \alpha\bar{\alpha},$$

y si hacemos $\alpha = (Qx) \cdot y$ queda $0 \leq -|(Qx) \cdot y|^2$, luego $(Qx) \cdot y = 0$.

Esto prueba la existencia de la descomposición de a). La unicidad se debe a que $V \cap V^\perp = 0$. En definitiva, $H = V \oplus V^\perp$.

Ahora observamos que si $y \in V$ entonces

$$\|x - y\|^2 = \|Qx + (Px - y)\|^2 = \|Qx\|^2 + \|Px - y\|^2,$$

luego la mínima distancia entre x y un punto $y \in V$ se alcanza cuando $y = Px$. Esto prueba b). El apartado d) es el teorema de Pitágoras. La linealidad de P y Q es obvia. La continuidad se debe a que por d) tenemos $\|Px\| \leq \|x\|$, $\|Qx\| \leq \|x\|$, y basta aplicar el teorema 2.37. ■

Terminamos con un teorema que caracteriza las aplicaciones lineales continuas de un espacio de Hilbert en $\mathbb{K}$.

Teorema 2.44 *Sea H un espacio de Hilbert y $f : H \longrightarrow \mathbb{K}$ una aplicación lineal continua. Entonces existe un único $y \in H$ tal que $f(x) = x \cdot y$ para todo $x \in H$.*

Demostración: Si f es la aplicación nula tomamos $y = 0$. En otro caso sea V el núcleo de f, que será un subespacio cerrado propio de H. Por el teorema anterior $V^\perp \neq 0$, luego podemos tomar $z \in V^\perp$ con $\|z\| = 1$. Sea $u = f(x)z - f(z)x$. Como $f(u) = f(x)f(z) - f(z)f(x) = 0$, tenemos que $u \in V$, luego $u \cdot z = 0$. La definición de u implica

$$f(x) = f(x)(z \cdot z) = f(z)(x \cdot z).$$

Tomando $y = \overline{f(z)}\, z$ resulta $f(x) = x \cdot y$.

La unicidad es obvia, pues si dos puntos y, y' cumplen el teorema, entonces $y - y'$ es ortogonal a todo $x \in H$. ■

2.5 Aplicaciones a las series numéricas

La completitud de $\mathbb{K}$ tiene muchas consecuencias sobre las series numéricas. En primer lugar, la condición de Cauchy para una serie en $\mathbb{K}$ puede expresarse como sigue:

Teorema 2.45 *Una serie* $\sum_{n=0}^{\infty} a_n$ *en* $\mathbb{K}$ *es convergente si y sólo si para todo* $\epsilon > 0$ *existe un natural* n_0 *tal que si* $n_0 \leq m \leq p$, *entonces* $\left|\sum_{n=m}^{p} a_n\right| < \epsilon$.

DEMOSTRACIÓN: Se trata de la condición de Cauchy, pues $\sum_{n=m}^{p} a_n$ es la diferencia entre la suma parcial p-ésima menos la suma parcial $m-1$-sima, y su módulo es la distancia entre ambas. ■

De aquí se sigue un hecho importantísimo.

Teorema 2.46 *Sea* $\sum_{n=0}^{\infty} a_n$ *una serie en* $\mathbb{K}$. *Si la serie* $\sum_{n=0}^{\infty} |a_n|$ *es convergente, entonces la serie* $\sum_{n=0}^{\infty} a_n$ *también lo es.*

DEMOSTRACIÓN: Dado $\epsilon > 0$ existe un número natural n_0 de manera que si $n_0 \leq m \leq p$, entonces $\sum_{n=m}^{p} |a_n| < \epsilon$.

Ahora bien, $\left|\sum_{n=m}^{p} a_n\right| \leq \sum_{n=m}^{p} |a_n| < \epsilon$, luego la serie sin módulos también es de Cauchy, luego converge. ■

Definición 2.47 Una serie $\sum_{n=0}^{\infty} a_n$ en $\mathbb{K}$ es *absolutamente convergente* si la serie $\sum_{n=0}^{\infty} |a_n|$ es convergente.

Hemos probado que toda serie absolutamente convergente es convergente. Las series convergentes que no son absolutamente convergentes se llaman series *condicionalmente convergentes.*

Un ejemplo de serie condicionalmente convergente es

$$1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \frac{1}{5} - \frac{1}{6} + \frac{1}{7} - \frac{1}{8} + \cdots = 0,693147\ldots$$

La convergencia absoluta de una serie es esencial para ciertas cuestiones. Por ejemplo, una consecuencia inmediata de las propiedades de las sumas finitas y de los límites de sucesiones es que

$$\sum_{n=0}^{\infty} a_n + \sum_{n=0}^{\infty} b_n = \sum_{n=0}^{\infty} (a_n + b_n), \quad a\sum_{n=0}^{\infty} a_n = \sum_{n=0}^{\infty} aa_n,$$

entendiendo que si las series de la izquierda convergen, las de la derecha también lo hacen y se da la igualdad. Un resultado análogo para producto de series ya no es tan sencillo.

Definición 2.48 Sean $\sum\limits_{n=0}^{\infty} a_n$ y $\sum\limits_{n=0}^{\infty} b_n$ dos series en $\mathbb{K}$. Llamaremos *producto de Cauchy* de estas series a la serie $\sum\limits_{n=0}^{\infty}\left(\sum\limits_{k=0}^{n} a_k \cdot b_{n-k}\right)$.

La intención es que la serie que acabamos de definir converja al producto de las dos series de partida, pero esto no ocurre necesariamente si al menos una de ellas no converge absolutamente.

Teorema 2.49 *Si* $\sum\limits_{n=0}^{\infty} a_n$ *y* $\sum\limits_{n=0}^{\infty} b_n$ *son dos series convergentes en* $\mathbb{K}$ *al menos una de las cuales converge absolutamente, entonces*

$$\sum_{n=0}^{\infty}\left(\sum_{k=0}^{n} a_k \cdot b_{n-k}\right) = \left(\sum_{n=0}^{\infty} a_n\right)\left(\sum_{n=0}^{\infty} b_n\right).$$

DEMOSTRACIÓN: Supongamos que la serie que converge absolutamente es $\sum\limits_{n=0}^{\infty} a_n$ y definamos

$$A = \sum_{n=0}^{\infty} a_n,\ B = \sum_{n=0}^{\infty} b_n,\ c_n = \sum_{k=0}^{n} a_k \cdot b_{n-k},\ C_n = \sum_{k=0}^{n} c_k,$$

$$A_n = \sum_{k=0}^{n} a_k,\ B_n = \sum_{k=0}^{n} b_k,\ \beta_n = B_n - B.$$

Ahora,

$$\begin{aligned} C_n &= c_0 + \cdots + c_n = a_0b_0 + (a_0b_1 + a_1b_0) + \cdots + (a_0b_n + \cdots + a_nb_0) \\ &= a_0B_n + \cdots + a_nB_0 = a_0(B + \beta_n) + \cdots + a_n(B + \beta_0) \\ &= A_nB + (a_0\beta_n + \cdots + a_n\beta_0) \end{aligned}$$

El teorema quedará probado si vemos que $a_0\beta_n + \cdots + a_n\beta_0$ tiende a 0.

Sea $\epsilon > 0$. Sea $K = \sum\limits_{n=0}^{\infty} |a_n|$. Sea $M = \sup\big\{|\beta_n| \mid n \geq 0\big\}$ (la sucesión β_n tiende a 0, luego está acotada).

Existe un número natural n_0 tal que si $n \geq n_0$, entonces $|\beta_n| < \epsilon/2K$ y si $k \geq n_0$, entonces $\sum\limits_{k=n_0+1}^{q} |a_k| < \epsilon/2M$. En consecuencia, si $n \geq 2n_0$,

$$\begin{aligned} |a_0b_n + \cdots + a_nb_0| &\leq \sum_{k=0}^{n} |a_k\beta_{n-k}| = \sum_{k=0}^{n_0} |a_k\beta_{n-k}| + \sum_{k=n_0+1}^{n} |a_k\beta_{n-k}| \\ &< \frac{\epsilon}{2K}\sum_{k=0}^{n} |a_k| + M\sum_{k=n_0+1}^{n} |a_k| \leq \frac{\epsilon}{2K}K + \frac{\epsilon}{2M}M = \epsilon. \end{aligned}$$

■

Si ninguna de las series converge absolutamente el resultado no tiene por qué cumplirse.

Ejemplo Consideremos la serie

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{\sqrt{n+1}}.$$

La serie converge por el criterio de Leibniz. El producto de Cauchy de esta serie por sí misma tiene término general

$$c_n = (-1)^n \sum_{k=0}^{n} \frac{1}{\sqrt{(n-k+1)(k+1)}}.$$

Cuando $n \geq k$ tenemos

$$(n-k+1)(k+1) = \left(\frac{n}{2}+1\right)^2 - \left(\frac{n}{2}-k\right)^2 \leq \left(\frac{n}{2}+1\right)^2,$$

luego

$$\frac{1}{\sqrt{(n-k+1)(k+1)}} \geq \frac{1}{\frac{n}{2}+1} = \frac{2}{n+2}.$$

Por consiguiente

$$|c_n| \geq \sum_{k=0}^{n} \frac{2}{n+2} = \frac{2(n+1)}{n+1}.$$

Esta expresión converge a 2, luego c_n no converge a 0 y el producto de Cauchy no converge.

El teorema anterior prueba, pues, que la serie dada es condicionalmente convergente. ■

Otro punto en el que la convergencia absoluta resulta crucial es en el de la reordenación de los términos de una serie.

Dada una serie convergente $\sum\limits_{n=0}^{\infty} a_n$ y una biyección $\sigma : \mathbb{N} \longrightarrow \mathbb{N}$, podemos considerar la serie $\sum\limits_{n=0}^{\infty} a_{\sigma(n)}$ y estudiar su convergencia. De nuevo el resultado natural exige que la serie converja absolutamente:

Teorema 2.50 *Si* $\sum\limits_{n=0}^{\infty} a_n$ *es una serie absolutamente convergente y* $\sigma : \mathbb{N} \longrightarrow \mathbb{N}$ *es una aplicación biyectiva, entonces la serie* $\sum\limits_{n=0}^{\infty} a_{\sigma(n)}$ *es absolutamente convergente y tiene la misma suma.*

Demostración: Una serie es absolutamente convergente si y sólo si las sumas parciales de sus módulos forman un conjunto acotado. Toda suma parcial de los módulos de la reordenación está mayorada por una suma parcial de los

módulos de la serie original (tomando los sumandos necesarios para incluir todos los que aparecen en la suma dada). Por tanto las sumas parciales de los módulos de la reordenación están acotadas y la serie converge absolutamente.

Sea $\epsilon > 0$. Existe un número natural n_0 tal que si $n \geq n_0$

$$\left|\sum_{k=0}^{n} a_k - \sum_{k=0}^{\infty} a_k\right| = \left|\sum_{k=n}^{\infty} a_k\right| \leq \sum_{k=n}^{\infty} |a_k| = \sum_{k=0}^{\infty} |a_k| - \sum_{k=0}^{n} |a_k| < \frac{\epsilon}{2}.$$

Sea $m_0 \geq n_0$ tal que $\{0, 1, \ldots, n_0\} \subset \{\sigma(0), \sigma(1), \ldots, \sigma(m_0)\}$. Entonces si $n \geq m_0$,

$$\begin{aligned}\left|\sum_{k=0}^{n} a_{\sigma(k)} - \sum_{k=0}^{\infty} a_k\right| &\leq \left|\sum_{k=0}^{n} a_{\sigma(k)} - \sum_{k=0}^{n_0} a_k\right| + \left|\sum_{k=0}^{n_0} a_k - \sum_{k=0}^{\infty} a_k\right| \\ &< \sum_{k=n_0+1}^{\infty} |a_k| + \frac{\epsilon}{2} < \epsilon.\end{aligned}$$

Por lo tanto $\sum\limits_{k=0}^{\infty} a_{\sigma(k)} = \sum\limits_{k=0}^{\infty} a_k$. ■

Como consecuencia, si I es cualquier conjunto infinito numerable y $\{a_i\}_{i\in I}$ es cualquier familia de elementos de $\mathbb{K}$ con la propiedad de que las sumas $\sum\limits_{i\in F} |a_i|$, con $F \subset I$ finito estén acotadas, tiene sentido la expresión $\sum\limits_{i\in I} a_i$, definida como la suma de la serie determinada por cualquier ordenación del conjunto I, y es un número independiente de la ordenación elegida. Obviamente la expresión $\sum\limits_{i\in I} a_i$ tiene también sentido cuando I es un conjunto finito.

Observar que si $\epsilon > 0$, existe un $F_0 \subset I$ finito tal que para todo $F_0 \subset F \subset I$, se cumple $\left|\sum\limits_{i\in I} a_i - \sum\limits_{i\in F} a_i\right| < \epsilon$. En efecto, basta considerar una ordenación de I y tomar como F_0 los primeros términos de la sucesión, de modo que el módulo de las colas con y sin módulos sea menor que $\epsilon/2$. Entonces

$$\left|\sum_{i\in I} a_i - \sum_{i\in F} a_i\right| \leq \left|\sum_{i\in I} a_i - \sum_{i\in F_0} a_i\right| + \left|\sum_{i\in F_0} a_i - \sum_{i\in F} a_i\right| < \frac{\epsilon}{2} + \sum_{i\in F\setminus F_0} |a_i| < \epsilon.$$

Las series absolutamente convergentes se pueden manipular exactamente igual que si fueran sumas finitas. El siguiente teorema justifica cualquier operación razonable entre ellas.

Teorema 2.51 *Sea $\{a_i\}_{i\in I}$ una familia de elementos de $\mathbb{K}$. Sea $I = \bigcup\limits_{n=0}^{\infty} I_n$ una división de I en partes disjuntas. Entonces $\sum\limits_{i\in I} a_i$ es (absolutamente) convergente si y sólo si lo son las series $\sum\limits_{i\in I_n} a_i$ y $\sum\limits_{n=0}^{\infty} |a_i|$. Además en tal caso*

$$\sum_{i\in I} a_i = \sum_{n=0}^{\infty} \sum_{i\in I_n} a_i.$$

Demostración: Si $\sum\limits_{i\in I} a_i$ es absolutamente convergente, sus sumas parciales en módulo están acotadas, pero toda suma parcial en módulo de cada $\sum\limits_{i\in I_n} |a_i|$ lo es también de la primera, luego éstas están acotadas, o sea, las series $\sum\limits_{i\in I_n} |a_i|$ convergen absolutamente. Dado cualquier natural k, tomamos para cada $n \leq k$ un conjunto finito $F_n \subset I_n$ tal que $\sum\limits_{i\in I_n} |a_i| - \sum\limits_{i\in F_n} |a_i| < 1/(k+1)$. Entonces

$$\sum_{n=0}^{k}\sum_{i\in I_n} |a_i| < \sum_{n=0}^{k}\sum_{i\in F_n} |a_i| + 1 \leq \sum_{i\in I} |a_i| + 1,$$

luego las sumas parciales están acotadas y así todas las series convergen absolutamente.

Supongamos ahora que las series $\sum\limits_{i\in I_n} |a_i|$ y $\sum\limits_{n=0}^{\infty}\sum\limits_{i\in I_n} |a_i|$ convergen absolutamente. Si $F \subset I$ es finito, para un cierto k suficientemente grande se cumple

$$\sum_{i\in F} |a_i| = \sum_{n=0}^{k}\sum_{i\in I_n\cap F} |a_i| \leq \sum_{n=0}^{k}\sum_{i\in I_n} |a_i| \leq \sum_{n=0}^{\infty}\sum_{i\in I_n} |a_i|,$$

luego las sumas parciales de $\sum\limits_{i\in I} |a_i|$ están acotadas y la serie converge absolutamente.

Ahora supongamos la convergencia de todas las series y probemos la igualdad de las sumas. Notemos que la serie

$$\sum_{n=0}^{\infty}\sum_{i\in I_n} a_i$$

es convergente porque es absolutamente convergente. Sea $\epsilon > 0$. Existe un número natural n_0 tal que

$$\left|\sum_{n=n_0+1}^{\infty}\sum_{i\in I_n} a_i\right| < \epsilon/4.$$

Para cada $n \leq n_0$ existe un conjunto finito $F_n \subset I_n$ tal que si $F_n \subset F \subset I_n$, entonces

$$\left|\sum_{i\in I_n} a_i - \sum_{i\in F_n} a_i\right| < \frac{\epsilon}{2(n_0+1)}.$$

Sea F un conjunto finito que contenga a todos los F_n y tal que

$$\left|\sum_{i\in I} a_i - \sum_{i\in F} a_i\right| < \epsilon/4.$$

Entonces

$$\begin{aligned}\left|\sum_{n=0}^{\infty}\sum_{i\in I_n} a_i - \sum_{i\in I} a_i\right| &\leq \left|\sum_{n=n_0+1}^{\infty}\sum_{i\in I_n} a_i\right| + \left|\sum_{n=0}^{n_0}\sum_{i\in I_n} a_i - \sum_{i\in F} a_i\right| \\ &+ \left|\sum_{i\in F} a_i - \sum_{i\in I} a_i\right| \\ &< \frac{\epsilon}{4} + \left|\sum_{n=0}^{n_0}\left(\sum_{i\in I_n} a_i - \sum_{i\in I_n\cap F} a_i\right)\right| + \frac{\epsilon}{4} + \frac{\epsilon}{2} + \frac{\epsilon}{4} = \epsilon.\end{aligned}$$

Por lo tanto ambas sumas coinciden. ■

Ejemplo Consideremos la serie condicionalmente convergente

$$S = \sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}.$$

Entonces

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{2n} = \frac{S}{2}.$$

Ahora consideremos la serie $\sum\limits_{n=1}^{\infty} a_n$ cuyos términos impares son ceros y sus términos pares son los de la serie anterior, es decir, la serie

$$0 + \frac{1}{2} + 0 - \frac{1}{4} + 0 + \frac{1}{6} + 0 - \frac{1}{8} + \cdots$$

Obviamente su suma es también $S/2$. La serie

$$\sum_{n=1}^{\infty}\left(\frac{(-1)^{n+1}}{n} + a_n\right)$$

converge a $3S/2$. Sus primeros términos son:

$$1 + 0 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + 0 + \frac{1}{7} - \frac{1}{4} + \frac{1}{9} + 0 + \cdots$$

Eliminando los ceros obtenemos la serie

$$1 + \frac{1}{3} - \frac{1}{2} + \frac{1}{5} + \frac{1}{7} - \frac{1}{4} + \frac{1}{9} + \frac{1}{11} - \frac{1}{6} \cdots$$

Vemos que se trata de una reordenación de la serie original, pero converge a $3S/2$. ■

2.6 Espacios de funciones

Uno de los éxitos de la topología consiste en que sus técnicas, desarrolladas en principio para estudiar espacios "geométricos" como $\mathbb{K}^n$, se aplican igualmente a objetos más abstractos, como son los conjuntos de funciones entre espacios topológicos. Las definiciones siguientes no corresponden en realidad a conceptos nuevos desde un punto de vista topológico:

Definición 2.52 Sea Y un espacio topológico y X un conjunto cualquiera. Una *sucesión funcional* de X en Y es una sucesión $\{f_n\}_{n=0}^{\infty}$ en el espacio Y^X de todas las aplicaciones de X en Y, es decir, para cada n se cumple $f_n : X \longrightarrow Y$.

Si Y es un espacio vectorial topológico (en especial si $Y = \mathbb{K}$) cada sucesión funcional define la correspondiente *serie funcional* $\sum\limits_{n=0}^{\infty} f_n$, es decir, la sucesión cuyos términos son las funciones $S_n = \sum\limits_{k=0}^{n} f_k : X \longrightarrow Y$.

Diremos que una sucesión funcional $\{f_n\}_{n=0}^{\infty}$ *converge puntualmente* a una función $f \in Y^X$ si para todo $x \in X$ se cumple $\lim\limits_n f_n(x) = f(x)$. En tal caso escribiremos $\lim\limits_n f_n = f$. Para series de funciones podemos definir de manera obvia la convergencia puntual absoluta y la convergencia puntual condicional.

En realidad no estamos introduciendo un nuevo concepto de convergencia. Notemos que Y^X es el producto cartesiano del espacio Y por sí mismo tantas veces como elementos tiene X, luego podemos considerarlo como espacio topológico con la topología producto. Las sucesiones convergen en esta topología si y sólo si convergen coordenada a coordenada, o sea, si y sólo si convergen puntualmente. Por ello a la topología producto en Y^X se la llama también *topología de la convergencia puntual.*

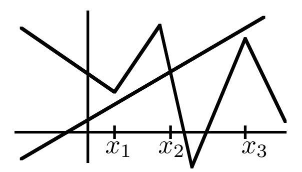
Sin embargo, la convergencia puntual no es la convergencia más natural que puede definirse sobre las sucesiones funcionales. De hecho presenta grandes inconvenientes.

Ejemplo Para cada $n \geq 1$ sea $f_n : \mathbb{R} \longrightarrow \mathbb{R}$ la función dada por

$$f_n(x) = \begin{cases} 0 & \text{si } x \leq 0 \\ nx & \text{si } 0 \leq x \leq 1/n \\ 1 & \text{si } 1/n \leq x \end{cases}$$

$1/n$

Si $x \leq 0$ entonces $f_n(x)$ es constante igual a 0 y si $x > 1$ entonces $f_n(x)$ es finalmente constante igual a 1, luego esta sucesión funcional converge puntualmente a la función f que muestra la figura de la izquierda. Tenemos, pues, una sucesión de funciones continuas cuyo límite puntual no es continuo. ■



En general el hecho de que una función sea límite puntual de una sucesión de funciones aporta muy poca información. La razón es que los entornos de la topología puntual son muy grandes. En efecto, en el caso de $\mathbb{R}^{\mathbb{R}}$ no es difícil ver que un entorno básico de una función f es un conjunto de la forma

$$\big\{g \in \mathbb{R}^{\mathbb{R}} \bigm| |g(x_i) - f(x_i)| < \epsilon,\ i = 1, \ldots, n\big\},$$

donde $\epsilon > 0$ y $x_1, \ldots, x_n \in \mathbb{R}$, es decir, si una sucesión funcional tiende a f, lo máximo que podemos garantizar tomando un índice grande es que los términos de la sucesión se parecerán a f en un número finito de puntos, pero dos funciones pueden parecerse en un número finito de puntos y ser muy diferentes.

Es mucho más natural considerar que dos funciones están próximas cuando distan menos de un ϵ en todos los puntos a la vez. Por ello, si Y es un espacio métrico, definimos la *topología de la convergencia uniforme* en Y^X como la que tiene por base de entornos abiertos de una función f a los conjuntos de la forma

$$B(f, \epsilon) = \{g \in Y^X \mid d\big(f(x), g(x)\big) < \epsilon \text{ para todo } x \in X\}.$$

De este modo, cuando una función g está en un entorno de f suficientemente pequeño, ambas funciones se parecen realmente. Es fácil comprobar que los conjuntos $B(f, \epsilon)$ cumplen las condiciones del teorema 1.14 y por tanto definen, según hemos dicho, una topología en Y^X.

Es inmediato comprobar que una sucesión funcional $\{f_n\}_{n=0}^{\infty}$ converge uniformemente (es decir, en la topología de la convergencia uniforme) a una función f si y sólo si para todo $\epsilon > 0$ existe un n_0 tal que si $n \geq n_0$, entonces $d\big(f_n(x), f(x)\big) < \epsilon$ para todo $x \in X$.

La diferencia, pues, entre la convergencia uniforme y la convergencia puntual es que cuando la convergencia es uniforme hay un n_0 a partir del cual todos los $f_n(x)$ distan de su límite menos de un ϵ dado, mientras que si la convergencia es puntual cada punto x puede requerir un n_0 mayor para acercarse a su límite en menos de ϵ, de manera que ningún n_0 sirva simultáneamente para todos los puntos.

Es obvio que si una sucesión funcional converge uniformemente a una función, también converge puntualmente a dicha función. Tanto la topología de la convergencia uniforme como la topología de la convergencia puntual son de Hausdorff, luego los límites son únicos.

Si en el espacio Y tomamos la distancia $d'(x, y) = \min\{1, d(x, y)\}$, los conjuntos $B(f, \epsilon)$ son los mismos cuando $\epsilon < 1$, luego d' induce la misma topología de convergencia uniforme en Y^X.

La ventaja de esta métrica es que no toma valores mayores que 1, por lo que podemos definir

$$d(f, g) = \sup\big\{d'\big(f(x), g(x)\big) \bigm| x \in X\big\},$$

y es claro que esta d es una distancia en Y^X para la cual $B(f, \epsilon)$ es simplemente la bola abierta de centro f y radio ϵ. Esto convierte a Y^X en un espacio métrico cuya topología es precisamente la de la convergencia uniforme.

Teorema 2.53 *Sea X un espacio topológico e Y un espacio métrico. Entonces el conjunto $C(X,Y)$ de las aplicaciones continuas de X en Y es cerrado en Y^X, cuando en éste consideramos la topología de la convergencia uniforme.*

Demostración: Sea $\{f_n\}_{n=0}^{\infty}$ una sucesión de aplicaciones continuas que converja uniformemente a una función f. Basta ver que f es continua.

Sea $\epsilon > 0$. Sea n_0 tal que si $n \geq n_0$ y $x \in X$, entonces $d\big(f_n(x), f(x)\big) < \epsilon/3$.

Sea $x_0 \in X$ (vamos a probar que f es continua en x_0). Como f_{n_0} es continua existe un entorno U de x_0 tal que si $x \in U$, entonces $d\big(f_{n_0}(x), f_{n_0}(x_0)\big) < \epsilon/3$. Por lo tanto, si $x \in U$, se cumple que

$$d\big(f(x), f(x_0)\big) \leq d\big(f(x), f_{n_0}(x)\big) + d\big(f_{n_0}(x), f_{n_0}(xo)\big) + d\big(f_{n_0}(x_0), f(x_0)\big) < \epsilon.$$

■

Éste es un primer ejemplo del buen comportamiento de la convergencia uniforme. Veamos otros:

Teorema 2.54 *Sea Y un espacio métrico completo y X un conjunto. Entonces Y^X es completo con la métrica inducida a partir de la métrica de Y. Por lo tanto si X es un espacio topológico, $C(X,Y)$ también es completo.*

Demostración: Ante todo notemos que si Y es completo con su métrica d, también lo es con la métrica d' que resulta de tomar el mínimo con 1, luego podemos suponer que la métrica en Y está acotada.

Sea $\{f_n\}_{n=0}^{\infty}$ una sucesión de Cauchy en Y^X. Esto significa que para todo $\epsilon > 0$ existe un n_0 tal que si m, $n \geq n_0$ y $x \in X$, entonces $d\big(f_m(x), f_n(x)\big) < \epsilon$.

En particular esto implica que cada sucesión $\{f_n(x)\}_{n=0}^{\infty}$ es de Cauchy en Y, luego converge a un cierto punto $f(x)$. Con esto tenemos una función $f \in Y^X$ a la cual $\{f_n\}_{n=0}^{\infty}$ converge puntualmente. Basta ver que también converge uniformemente.

Sea $\epsilon > 0$ y tomemos un natural n_0 como antes. Así, si $n_0 \leq n \leq m$, se cumple $d\big(f_m(x), f_n(x)\big) < \epsilon$, luego $f_m(x)$ está en la bola cerrada de centro $f_n(x)$ y radio ϵ, luego el límite $f(x)$ estará en esta misma bola, o sea, se cumplirá $d\big(f(x), f_n(x)\big) \leq \epsilon$, y esto para todo $n \geq n_0$ y todo $x \in X$. Esto significa que la sucesión converge uniformemente a f. La completitud de $C(X,Y)$ se sigue de que es un cerrado. ■

En general no podemos convertir a Y^X en un espacio normado aunque Y lo sea. El problema es que no podemos transformar la norma en una norma acotada. Lo único que podemos hacer es definir $(Y^X)^*$ como el espacio de las funciones acotadas de X en Y, es decir, las funciones f tales que $f[X]$ está acotado. En este conjunto podemos definir la *norma supremo* dada por $\|f\|_\infty = \sup\big\{\|f(x)\| \bigm| x \in X\big\}$, que obviamente genera las bolas que definen la topología de la convergencia uniforme (restringida a $(Y^X)^*$). Así pues, si Y es un espacio normado, $(Y^X)^*$ también lo es, y como es fácil ver que el límite uniforme de funciones acotadas está acotado, resulta que $(Y^X)^*$ es cerrado en Y^X, luego si Y es un espacio de Banach, $(Y^X)^*$ también lo es.

Si X es un espacio topológico e Y es un espacio normado, definimos $C^*(X,Y)$ como el espacio de las funciones continuas y acotadas de X en Y. Obviamente se trata de la intersección de dos cerrados, luego es cerrado, es un espacio normado y si Y es un espacio de Banach, $C^*(X,Y)$ también lo es.

En particular $C^*(X,\mathbb{K})$ es un espacio de Banach. En general no podemos dotar a $C(X,\mathbb{K})$ de estructura de espacio normado. De hecho no es un espacio vectorial topológico porque no es conexo. En efecto, es fácil ver que $C^*(X,\mathbb{K})$ es abierto y cerrado en $C(X,\mathbb{K})$. La única excepción es precisamente cuando $C(X,\mathbb{K}) = C^*(X,\mathbb{K})$. Esto ocurre por ejemplo si X es un compacto. Es decir, si X es compacto, entonces $C(X,\mathbb{K})$ es un espacio de Banach con la norma supremo y la topología es la de la convergencia uniforme.

Sea $L(E,F)$ el conjunto de las aplicaciones lineales continuas entre dos espacios normados. Teniendo en cuenta 2.37, la aplicación $L(E,F) \longrightarrow C^*(\overline{B_1(0)},F)$ definida por restricción es claramente lineal e inyectiva (pues $\overline{B_1(0)}$ contiene una base de E). Si transportamos la norma de este segundo espacio al primero obtenemos que, para cada aplicación lineal y continua $f : E \longrightarrow F$, su norma es el supremo de f en $\overline{B_1(0)}$, y por lo tanto cumple $\|f(v)\| \leq \|f\|\,\|v\|$ para todo $v \in E$.

Ejercicio: Probar que $L(E,F)$ es un subespacio cerrado de $C^*(\overline{B_1(0)},F)$. Por consiguiente, si F es un espacio de Banach, $L(E,F)$ también lo es.

Terminamos con un resultado importante sobre convergencia de series funcionales. Previamente notemos lo siguiente: si una serie $\sum\limits_{n=0}^{\infty} f_n$ converge absoluta y uniformemente en un conjunto X, es decir, si la serie $\sum\limits_{n=0}^{\infty} |f_n|$ converge uniformemente, entonces $\sum\limits_{n=0}^{\infty} f_n$ converge uniformemente. La prueba es la misma que la de 2.46.

Teorema 2.55 (Criterio de Mayoración de Weierstrass) *Sea $\sum\limits_{n=0}^{\infty} f_n$ una serie funcional en un espacio X y $\{M_n\}_{n=0}^{\infty}$ una sucesión en el intervalo $[0,+\infty[$ tal que para todo natural n y todo $x \in X$ se cumpla $|f_n(x)| \leq M_n$. Si la serie $\sum\limits_{n=0}^{\infty} M_n$ es convergente, entonces la serie f_n es absoluta y uniformemente convergente en X.*

DEMOSTRACIÓN: La serie M_n es de Cauchy, luego dado $\epsilon > 0$ existe un n_0 tal que si $n_0 \leq m \leq p$, entonces $\sum\limits_{n=m}^{p} M_n < \epsilon$. Así

$$\left|\sum_{n=m}^{p} |f_n(x)|\right| = \sum_{n=m}^{p} |f_n(x)| \leq \sum_{n=m}^{p} M_n < \epsilon$$

para todo $x \in X$. Esto significa que la serie $\sum\limits_{n=0}^{\infty} |f_n|$ es de Cauchy en $C(X,\mathbb{K})$,

luego (uniformemente) convergente, luego $\sum_{n=0}^{\infty} f_n$ es absoluta y uniformemente convergente. ■

2.7 Apéndice: El teorema de Baire

Terminamos el capítulo con un resultado que no nos va a hacer falta más adelante, pero que es útil en algunos contextos más avanzados. Necesitamos algunos resultados previos:

Definición 2.56 Si M es un espacio métrico y $C \subset M$, se define el *diámetro* de C como

$$d(C) = \sup\{d(x, y) \mid x,\, y \in C\} \in [0, +\infty].$$

Es fácil calcular el diámetro de una bola abierta:

$$d(B_r(x)) = 2r.$$

También se comprueba sin dificultad que el diámetro de un conjunto coincide con el de su clausura. Por último, necesitamos el siguiente hecho elemental:

Teorema 2.57 *Sea M un espacio métrico completo. Toda familia decreciente $\{C_n\}_n$ de cerrados en M no vacíos tal que $\lim_n d(C_n) = 0$ tiene intersección no vacía.*

Demostración: Para cada n tomamos $x_n \in C_n$. Como $\lim_n d(C_n) = 0$, es claro que la sucesión $(x_n)_n$ es de Cauchy. Su límite x está en cada C_n por ser éste cerrado y $\{C_n\}_n$ decreciente. ■

Teorema 2.58 (Teorema de Baire) *En un espacio métrico completo, la intersección de una familia numerable de abiertos densos es un conjunto denso.*

Demostración: Sea M un espacio métrico completo y $G = \bigcap_{n=1}^{\infty} G_n$ una intersección numerable de abiertos G_n densos en M. Basta probar que G corta a toda bola abierta $B_r(x)$. Como G_1 es denso en M existe $x_1 \in G_1 \cap B_r(x)$. Como $G_1 \cap B_r(x)$ es abierto, existe un $r_1 > 0$, que podemos tomar menor que $r/2$, tal que $\overline{B}_{r_1}(x_1) \subset G_1 \cap B_r(x)$.

Inductivamente podemos construir una sucesión $\{x_n\}_n$ de puntos de M y una sucesión $\{r_n\}_n$ de números reales positivos de modo que

$$\overline{B}_{r_n}(x_n) \subset G_n \cap B_{r_{n-1}}(x_{n-1}) \qquad (2.2)$$

y $r_n < r/n$ para todo n.

Por el teorema anterior, $\bigcap\limits_{n=1}^{\infty} \overline{B}_{r_n}(x_n) \neq \varnothing$, luego (2.2) implica que

$$G \cap B_r(x) \supset \bigcap_{n=1}^{\infty} (G_n \cap B_{r_{n-1}}(x_{n-1})) \neq \varnothing.$$

■

Sin más que tener en cuenta que el complementario de un conjunto denso es un conjunto con interior vacío tenemos una forma equivalente del teorema de Baire:

Teorema 2.59 (Teorema de Baire) *En un espacio métrico completo, toda unión numerable de cerrados de interior vacío tiene interior vacío.*

Conviene observar que la tesis del teorema de Baire (en cualquiera de sus dos formas equivalentes) se cumple también sobre espacios topológicos localmente compactos, no necesariamente metrizables. La prueba es, de hecho, más sencilla, y se obtiene sustituyendo el teorema 2.57 por el hecho de que la intersección de una familia decreciente de compactos no vacíos es no vacía.

Aunque, según hemos indicado, no necesitaremos el teorema de Baire, vamos a tratar de explicar su interés. Para ello necesitamos algunas definiciones:

Definición 2.60 Sea X un espacio topológico. Un subconjunto A de X es *diseminado* si $X \setminus A$ contiene un abierto denso o, equivalentemente, si A está contenido en un cerrado de interior vacío o, también, si $\operatorname{int} \overline{A} = \varnothing$.

Informalmente, la idea es que un abierto denso (y cualquier conjunto que lo contenga) es un conjunto "muy grande" desde el punto de vista topológico, pues todo abierto contiene un abierto contenido en tal conjunto; los conjuntos diseminados son topológicamente "muy pequeños". Como todo conjunto que contenga a un conjunto que contenga a un abierto denso contiene un abierto denso, tomando complementarios obtenemos que los subconjuntos de los conjuntos diseminados son diseminados. Esta noción de conjunto diseminado resulta ser muy restrictiva, esencialmente a causa de que no se conserva por uniones numerables, por ello se definen los conjuntos de primera categoría:

Definición 2.61 Un subconjunto A de un espacio topológico X es de *primera categoría* si es unión numerable de conjuntos diseminados. A es de *segunda categoría* si no es de primera categoría.

Es evidente que toda unión numerable de conjuntos de primera categoría es de primera categoría. Así, los conjuntos de primera categoría son conjuntos topológicamente "pequeños", aunque no necesariamente "muy pequeños", mientras que los conjuntos de segunda categoría son los topológicamente "grandes".

No obstante, estas nociones no sirven de nada sin el teorema de Baire, que puede enunciarse en una tercera forma equivalente:

Teorema 2.62 (Teorema de Baire) *En un espacio métrico completo, los conjuntos de primera categoría tienen interior vacío.*

DEMOSTRACIÓN: Consideremos un conjunto C de primera categoría. Entonces

$$C = \bigcup_n A_n \subset \bigcup_n \overline{A}_n = C',$$

donde los conjuntos A_n son diseminados, luego sus clausuras son cerrados de interior vacío, luego C' tiene interior vacío (por la versión que ya hemos probado del teorema de Baire) y C también. ■

Así pues, si probamos que un conjunto C es "pequeño", en el sentido de que es de primera categoría, el teorema de Baire nos da que todo abierto va a contener puntos que no están en C. Éste es esencialmente el interés del teorema de Baire.

Terminaremos con una aplicación del teorema de Baire, que no es de las más típicas, pero tal vez la más sencilla:

Ejemplo *$\mathbb{Q}$ no puede expresarse como una intersección numerable de abiertos de $\mathbb{R}$.*

En efecto, en tal caso sería una intersección numerable de abiertos densos, luego $\mathbb{R} \setminus \mathbb{Q}$ sería una unión numerable de cerrados de interior vacío, al igual que $\mathbb{R} = \mathbb{Q} \cup (\mathbb{R} \setminus \mathbb{Q})$. El teorema de Baire nos daría entonces que $\mathbb{R}$ tiene interior vacío (en sí mismo), lo cual es absurdo. ■

Los conjuntos que pueden expresarse como intersección numerable de abiertos se llaman conjuntos G_δ. El ejemplo anterior, junto con el teorema siguiente, muestra que no puede existir una función $f : \mathbb{R} \longrightarrow \mathbb{R}$ continua en los puntos $\mathbb{Q}$ y discontinua en los de $\mathbb{R} \setminus \mathbb{Q}$. (Si el lector cree que esto es evidente, debería pensar en el ejercicio que sigue al teorema.)

Teorema 2.63 *Sea M un espacio métrico completo. El conjunto de puntos de continuidad de toda función $f : M \longrightarrow \mathbb{R}$ es un G_δ.*

DEMOSTRACIÓN: Para cada natural no nulo n definimos

$$G_n = \{x \in M \mid \text{ existe } \delta > 0 \text{ tal que } \sup_{y \in B_\delta(x)} f(y) - \inf_{y \in B_\delta(x)} f(y) < 1/n\}.$$

Claramente los conjuntos G_n son abiertos. Basta probar que el conjunto de puntos de continuidad de f es $G = \bigcap_{n=1}^{\infty} G_n$.

Si f es continua en x, dado un $n > 0$ existe un $\delta > 0$ tal que si $y \in B_\delta(x)$ entonces $|f(x) - f(y)| < 1/(4n)$. Por lo tanto, si y, $y' \in B_\delta(x)$ se cumple que $|f(y) - f(y')| < 1/(2n)$ y, tomando el supremo en y y el ínfimo en y', concluimos que $x \in G_n$.

Recíprocamente, supongamos que $x \in G$. Dado $\epsilon > 0$ tomamos n tal que $1/n < \epsilon$. Como $x \in G_n$, existe $\delta > 0$ tal que

$$\sup_{y \in B_\delta(x)} f(y) - \inf_{y \in B_\delta(x)} f(y) < 1/n.$$

Entonces si $y \in B_\delta(x)$ se tiene que

$$|f(x) - f(y)| \leq \sup_{y \in B_\delta(x)} f(y) - \inf_{y \in B_\delta(x)} f(y) < 1/n < \epsilon,$$

luego f es continua en x. ■

Ejercicio: Consideremos la función $f : \mathbb{R} \longrightarrow \mathbb{R}$ dada por

$$f(x) = \begin{cases} 1/q & \text{si } x = p/q \text{ con } p,\, q \in \mathbb{Z},\ (p,q) = 1,\ q > 0, \\ 0 & \text{en caso contrario.} \end{cases}$$

Demostrar que $\lim_{x \to x_0} f(x) = 0$ para todo $x_0 \in \mathbb{R}$. Deducir que f es continua en $\mathbb{R} \setminus \mathbb{Q}$ y discontinua en $\mathbb{Q}$.

Capítulo III

Cálculo diferencial de una variable

En este capítulo estudiaremos una de las ideas más importantes y fructíferas que posee la matemática actual. Su núcleo está en la observación de que muchas curvas se parecen localmente a rectas. Por ejemplo, visto suficientemente de cerca, un arco de circunferencia es indistinguible de un segmento de recta. Una prueba de ello está en el horizonte que separa el cielo del mar en un día despejado. Se trata de un arco de circunferencia, pero ¿se ve que es así?

La razón por la que el horizonte parece recto no es que la Tierra sea muy grande, sino que la vemos muy de cerca. Vista desde la Luna es claramente esférica y cualquier circunferencia al microscopio parece una recta. Lo mismo sucede con muchas curvas, que si las vemos muy de cerca parecen rectas. Pero esto no es cierto para todas. Pensemos en la curva de la figura.

Vista de cerca podría pasar por una recta en un entorno de cualquiera de sus puntos excepto el señalado con el círculo, donde tiene un "pico". Por más que nos acerquemos nunca dejaremos de ver ese pico que delatará que no se trata de una recta. Vamos a estudiar las curvas que localmente se parecen a rectas. Pero ¿qué es exactamente parecerse a una recta?

3.1 Derivación

Consideremos una función f definida en un entorno de un punto $a \in \mathbb{R}$ y con imagen en $\mathbb{R}$. Supongamos que su gráfica se parece mucho a una recta en un entorno del punto. La pregunta es ¿a qué recta se parece? Por lo pronto a una que pasa por el punto $\big(a, f(a)\big)$. Las rectas que pasan por dicho punto (a excepción de la vertical, que no nos va a interesar) son de la forma $r(x) = m(x-a) + f(a)$, donde $m \in \mathbb{R}$.

La interpretación geométrica de m es sencilla. En general, si tenemos una recta $r(x) = mx + n$, en un punto a tomará el valor $r(a) = ma + n$. Si nos trasladamos a un punto $a+h$ con $h \neq 0$ obtendremos $r(a+h) = ma+mh+n$. El incremento que ha experimentado la función es $r(a+h) - r(a) = mh$, y si lo dividimos por el desplazamiento h obtenemos, independientemente de cuál sea h, el valor m. Es decir,

$$m = \frac{r(a+h) - r(a)}{h}.$$

Geométricamente esto no es sino el teorema de Tales. En definitiva, m expresa lo que aumenta la función por unidad de avance: si nos desplazamos $h = 1$ unidad, la recta aumenta en m unidades, si avanzamos $h = 2$ unidades, la recta aumenta $2m$, etc. Por lo tanto, si el valor de m es grande la recta subirá muy rápidamente, será una recta muy empinada. Si $m = 0$ la recta no sube, es constante. Si m es negativo la recta baja, más rápidamente cuanto mayor sea m en módulo. El número m se llama pendiente de la recta. Una recta viene determinada por dos de sus puntos o bien por uno de sus puntos y su pendiente (pues conocido un punto (a, b) y la pendiente m conocemos más puntos: $(a + 1, b + m)$, por ejemplo).

Volviendo a nuestro problema, tenemos la recta $r(x) = m(x-a)+f(a)$, cuya pendiente es m. Nos falta determinar m para que sea la recta que se parece a f. Consideremos la expresión

$$m(h) = \frac{f(a+h) - f(a)}{h}. \tag{3.1}$$

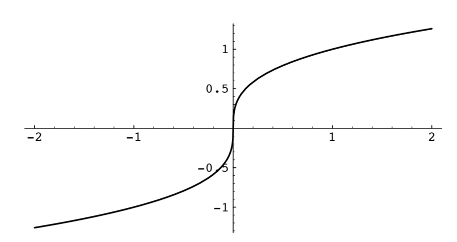
Esto no es una constante (salvo que f sea una recta), pero si ciertamente f se parece a una recta r, esta expresión debería parecerse a la pendiente de r. Si f se parece más a r cuanto más de cerca la miramos, esto es, cuando consideramos puntos más cercanos al punto a, el valor $m(h)$ debería parecerse más a la pendiente de r cuanto menor es h. Por ello definimos:

Definición 3.1 Sea $f : A \longrightarrow \mathbb{R}$ y a un punto interior de A. Diremos que f es *derivable* en a si existe (en $\mathbb{R}$)

$$f'(a) = \lim_{h\to 0} \frac{f(a+h) - f(a)}{h}.$$

Cuando esto sucede, a la recta $r(x) = f'(a)(x - a) + f(a)$ se le llama *recta tangente* a f en el punto $\big(a, f(a)\big)$ (o para abreviar, en el punto a). El número $f'(a)$ es la *derivada* de f en el punto a.

Según hemos dicho, una función es derivable en un punto cuando su gráfica se confunde en un entorno de dicho punto con la de una recta, la recta tangente a la función en el punto. Esto no es exacto, pues en realidad hay funciones que se parecen a rectas en los alrededores de un punto y pese a ello no son derivables. Esto ocurre cuando la recta tangente es vertical, con lo que su pendiente es infinita y no existe (en $\mathbb{R}$) el límite que define la derivada. Un ejemplo lo proporciona la función $\sqrt[3]{x}$ en $x = 0$.



Observamos que el eje vertical es tangente a la función en 0, pese a lo cual, según veremos, la función no es derivable en 0.

Diremos que una función es *derivable* en un abierto A si es derivable en todos los puntos de A. Una función es *derivable* si su dominio es un abierto y es derivable en todos sus puntos.

Si $f : A \longrightarrow \mathbb{R}$ es derivable, tenemos definida otra función $f' : A \longrightarrow \mathbb{R}$ que a cada punto $a \in A$ le asigna su derivada $f'(a)$. A esta función la llamamos (función) *derivada* de f en A.

Teniendo en cuenta la motivación que hemos dado para el concepto de derivada, es claro que toda recta no vertical, $f(x) = mx + n$ es derivable en $\mathbb{R}$ y su derivada es su pendiente, o sea, m. La razón es que, según hemos visto, el cociente (3.1) es en este caso constante igual a m, luego el límite cuando h tiende a 0 es igualmente m. En particular, la derivada de una función constante, $f(x) = a$, es $f'(x) = 0$.

Ejemplo Calculemos la derivada de la función $f(x) = x^2$.

$$f'(x) = \lim_{h \to 0} \frac{(x+h)^2 - x^2}{h} = \lim_{h \to 0} \frac{x^2 + 2xh + h^2 - x^2}{h} = \lim_{h \to 0} 2x + h = 2x.$$

■

Enseguida veremos que es muy fácil reconocer las funciones derivables así como calcular sus derivadas. Primero demostremos un hecho básico. Obviamente, lo primero que ha de hacer una función para parecerse a una recta es ser continua.

Teorema 3.2 *Si una función es derivable en un punto a, entonces es continua en a.*

DEMOSTRACIÓN: Sea $f : A \longrightarrow \mathbb{R}$ derivable en a. Entonces existe

$$f'(a) = \lim_{h \to 0} \frac{f(a+h) - f(a)}{h}.$$

Como $\lim\limits_{h \to 0} h = 0$, multiplicando obtenemos que

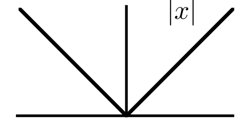
$$\lim_{h \to 0} (f(a+h) - f(a)) = f'(a)0 = 0.$$

Por lo tanto $\lim_{h\to 0} f(a+h) = f(a)$. Teniendo en cuenta la definición de límite, es fácil ver que esto equivale a que $\lim_{x\to a} f(x) = f(a)$. Esto significa que f es continua en a. ■

En particular, las funciones derivables son continuas, pero no toda función continua es derivable.

Ejemplos Ya hemos dicho que $\sqrt[3]{x}$ no es derivable en 0. Es fácil probarlo. La derivada en 0 sería

$$\lim_{h\to 0} \frac{\sqrt[3]{h}-0}{h} = \lim_{h\to 0} \frac{1}{\sqrt[3]{h^2}} = +\infty.$$



Aquí la razón es que la pendiente de la función se vuelve infinita en 0. Otra causa de no derivabilidad (a pesar de la continuidad) es que la función forme un "pico". Por ejemplo $f(x) = |x|$ en $x = 0$. La derivada sería

$$\lim_{h\to 0} \frac{|h|-0}{h} = \lim_{h\to 0} \operatorname{sig} h,$$

pero es claro que el límite por la izquierda es -1 y el límite por la derecha es $+1$, luego no existe tal límite. ■

3.2 Cálculo de derivadas

El teorema siguiente recoge las propiedades básicas que nos permiten derivar las funciones más simples:

Teorema 3.3 *Sean f, $g : A \longrightarrow \mathbb{R}$ funciones derivables en un punto $a \in A$ y $\alpha \in \mathbb{R}$.*

a) $f+g$ es derivable en a y $(f+g)'(a) = f'(a) + g'(a)$.

b) αf es derivable en a y $(\alpha f)'(a) = \alpha f'(a)$.

c) fg es derivable en a y $(fg)'(a) = f'(a)g(a) + f(a)g'(a)$.

d) Si $g(a) \neq 0$, f/g es derivable en a y

$$(f/g)'(a) = \frac{f'(a)g(a) - f(a)g'(a)}{g^2(a)}.$$

En particular, las funciones derivables en A forman una subálgebra de $C(A)$.

DEMOSTRACIÓN: Las propiedades a) y b) son muy sencillas. Veamos c).

$$\begin{aligned}(fg)'(a) &= \lim_{h\to 0}\frac{g(a+h)g(a+h)-f(a)g(a)}{h}\\ &= \lim_{h\to 0}\frac{f(a+h)g(a+h)-f(a)g(a+h)+f(a)g(a+h)-f(a)g(a)}{h}\\ &= \lim_{h\to 0}\left(\frac{f(a+h)-f(a)}{h}\,g(a+h)+f(a)\,\frac{g(a+h)-g(a)}{h}\right)\\ &= f'(a)g(a)+f(a)g'(a),\end{aligned}$$

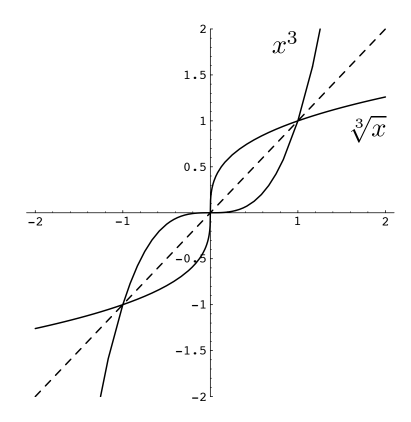
donde hemos usado la continuidad de g en a al afirmar que $\lim_{h\to 0} g(a+h) = g(a)$.

La prueba de d) es similar. ■

Aplicando inductivamente la propiedad c) se obtiene que $(x^n)' = nx^{n-1}$, para todo natural $n \geq 1$. Aplicando ahora d) resulta que esto es cierto para todo entero $n \neq 0$. En particular todos los polinomios y fracciones algebraicas son derivables en sus dominios.

Por ejemplo, la derivada de $3x^4-2x^3+x^2+5x-3$ es igual a $12x^3-6x^2+2x+5$. Observar que la derivada de un polinomio coincide con su derivada formal en el sentido algebraico.

La derivación de las raíces se seguirá un resultado general sobre funciones inversas. Por ejemplo, la función $\sqrt[3]{x}$ es la inversa de la función x^3. Esto significa que un punto (x, y) está en la gráfica de $\sqrt[3]{x}$ si y sólo si el punto (y, x) está en la de x^3. La gráfica de una se obtiene de la de la otra cambiando x por y. Geométricamente esto equivale a girar la gráfica respecto a la diagonal:



Es claro geométricamente que si una función tiene tangente en el punto (x, y), su inversa tendrá tangente en el punto (y, x), y que la recta tangente a la inversa resultará de girar respecto a la diagonal la recta tangente a la función original. Ahora bien, ¿qué relación hay entre la pendiente de una recta y la pendiente de la recta que resulta de girarla respecto a la diagonal?. Si una recta es $y = mx + n$ (con pendiente m), girar respecto a la diagonal es cambiar y por x, o sea, pasar a $x = my + n$. Despejando y obtenemos $y = (1/m)x - n/m$.

Por lo tanto la pendiente es $1/m$. Si la recta original tiene pendiente 0 (es horizontal), la recta girada es vertical, tiene pendiente infinita.

Por lo tanto es de esperar que si una función biyectiva f cumple $f(a) = b$ y tiene derivada $m \neq 0$ en a, entonces su inversa tiene derivada $1/m$ en b. Antes de probarlo analíticamente damos un sencillo resultado técnico.

Definición 3.4 Sea A un intervalo y $f : A \longrightarrow \mathbb{R}$, diremos que f es *creciente* en A si cuando $x < y$ son dos puntos de A, se cumple $f(x) \leq f(y)$. Si de hecho se cumple $f(x) < f(y)$ diremos que f es *estrictamente creciente* en A. Se dice que f es *decreciente* en A si cuando $x < y$ son puntos de A, se cumple $f(y) \leq f(x)$. Si se cumple $f(y) < f(x)$ se dice que es *estrictamente decreciente* en A. La función f es (estrictamente) monótona en A si es (estrictamente) creciente o decreciente en A.

Por ejemplo, la función x^3 es estrictamente creciente en $\mathbb{R}$.

Teorema 3.5 *Sea A un intervalo y $f : A \longrightarrow \mathbb{R}$ una función inyectiva y continua. Entonces f es estrictamente monótona en A.*

DEMOSTRACIÓN: Sean $a < b$ dos puntos cualesquiera de A. Supongamos que $f(a) < f(b)$. Entonces todo $a < x < b$ ha de cumplir $f(a) < f(x) < f(b)$, pues si, por ejemplo, $f(x) < f(a) < f(b)$, por el teorema de los valores intermedios, en el intervalo $]x, b[$ habría un punto cuya imagen sería $f(a)$, y f no sería inyectiva.

De aquí se sigue que f es creciente en $[a, b]$, pues si $a \leq x < y \leq b$, hemos visto que $f(a) < f(x) < f(b)$, y aplicando lo mismo a los puntos x, b, resulta que $f(x) < f(y) < f(b)$. Igualmente, de $f(b) < f(a)$ llegaríamos a que f es decreciente en $[a, b]$. Por lo tanto f es monótona en cualquier intervalo $[a, b]$ contenido en A.

Pero si f no fuera monótona en A existirían puntos $u < v$ y $r < s$ tales que $f(u) < f(v)$ y $f(s) < f(r)$. Tomando el mínimo y el máximo de estos cuatro puntos obtendríamos los extremos de un intervalo en el que f no sería monótona. ■

Teorema 3.6 (Teorema de la función inversa) *Sea A un intervalo abierto y $f : A \longrightarrow \mathbb{R}$ una función inyectiva y derivable en A tal que f' no se anule en ningún punto de A. Entonces*

a) $B = f[A]$ es un intervalo abierto.

b) La función inversa $g = f^{-1} : B \longrightarrow A$ es derivable en B.

c) Para todo $a \in A$, si $f(a) = b$, se cumple que $g'(b) = 1/f'(a)$.

DEMOSTRACIÓN: Por el teorema anterior sabemos que f es estrictamente monótona. Digamos que es monótona creciente. Si $a < b$ son dos puntos de A, por conexión $f\big[]a, b[\big]$ ha de ser un intervalo, y de la monotonía se sigue fácilmente que $f\big[]a, b[\big] =]f(a), f(b)[$.

Dado $a \in A$, podemos tomar un $\epsilon > 0$ tal que $[a-\epsilon, a+\epsilon] \subset A$, con lo que

$$f(a) \in \left]f(a-\epsilon), f(a-\epsilon)\right[= f\left[\left]a-\epsilon, a+\epsilon\right[\right] \subset B.$$

Así pues B es un entorno de $f(a)$ para todo $a \in A$, o sea, para todos los puntos de B. Por lo tanto B es abierto. Por conexión es un intervalo. Además hemos visto que f envía abiertos básicos $]a,b[$ a abiertos básicos, y esto significa que g es continua.

Sea ahora $f(a) = b$. Por la monotonía, si $h \neq 0$, entonces $g(b+h) \neq g(b)$. Sea $k = g(b+h) - g(b) \neq 0$. Así $g(b+h) = k + a$, luego $b + h = f(k+a)$, y $h = f(k+a) - f(a)$. Por lo tanto

$$\frac{g(b+h)-g(b)}{h} = \frac{1}{\frac{f(a+k)-f(a)}{k}}$$

Ahora, k es una función de h y, como g es continua, $\lim\limits_{h\to 0} k(h) = 0$. La derivabilidad de f en el punto a nos da que

$$g'(b) = \lim_{h\to 0} \frac{g(b+h)-g(b)}{h} = \frac{1}{f'(a)}.$$

■

Ejemplo Sea n un número natural no nulo y consideremos $f(x) = x^n$ definida en $]0,+\infty[$. Sabemos que es inyectiva y derivable. Su derivada es nx^{n-1}, que no se anula en $]0,+\infty[$. Por lo tanto su inversa, que es $g(x) = \sqrt[n]{x}$, es derivable en su dominio y si $y^n = x$ (con lo que $y = \sqrt[n]{x}$), entonces $g'(x) = 1/f'(y)$, o sea,

$$(\sqrt[n]{x})' = \frac{1}{ny^{n-1}} = \frac{1}{n\sqrt[n]{x}^{\,n-1}} = \frac{1}{n}x^{-1+1/n}.$$

Así pues, tenemos probado que la regla de derivación $x^r \mapsto rx^{r-1}$ es válida cuando r es entero o de la forma $1/n$, donde n es un número natural no nulo. Aplicando la regla del producto se concluye por inducción que vale de hecho para todo número racional r. ■

Con todo lo anterior, todavía no sabemos derivar funciones como $\sqrt{x^2+1}$. Con el teorema siguiente estaremos en condiciones de derivar cualquiera de las funciones que podemos construir a partir de polinomios y raíces.

Teorema 3.7 (regla de la cadena) *Sean $f : A \longrightarrow \mathbb{R}$ y $g : B \longrightarrow \mathbb{R}$. Sea un punto $a \in A$ tal que f sea derivable en a y g sea derivable en $f(a)$. Entonces la función compuesta $f \circ g$ es derivable en a y $(f \circ g)'(a) = g'(f(a))f'(a)$.*

DEMOSTRACIÓN: Notemos que B es un entorno de $f(a)$ y f es continua en a, luego $f^{-1}[B]$ es un entorno de a sobre el que está definida $f \circ g$.

Sea $b = f(a)$. Para $k \neq 0$, llamemos

$$G(k) = \frac{g(b+k)-g(b)}{k} - g'(b).$$

La función G está definida para los puntos k tales que $b+k \in B$. Como B es abierto, G está definida al menos para h en un intervalo $]-\epsilon, \epsilon[\setminus \{0\}$. Como g es derivable en b, existe $\lim\limits_{k \to 0} G(k) = 0$, luego si definimos $G(0) = 0$ tenemos que G es continua en un entorno de 0. Claramente además

$$g(b+k) - g(b) = \big(g'(b) + G(k)\big)k.$$

Ahora tomamos $h \neq 0$ tal que $a+h \in A$ y $k = f(a+h) - f(a)$, con lo que se cumple $f(a+h) = b+k$, luego $b+k \in B$ y está definido $G(k)$. Entonces

$$\begin{aligned} g\big(f(a+h)\big) - g\big(f(a)\big) &= \big(g'(f(a)) + G(k)\big)k \\ &= \big(g'(f(a)) + G(k)\big)\big(f(a+h) - f(a)\big). \end{aligned}$$

En consecuencia

$$\frac{(f \circ g)(a+h) - (f \circ g)(a)}{h} = \big(g'(f(a)) + G(f(a+h) - f(a))\big)\frac{f(a+h) - f(a)}{h}.$$

Usando la continuidad de f en a y la de G en 0, tomamos el límite cuando h tiende a 0 y queda que existe $(f \circ g)'(a) = g'(f(a))f'(a)$. ■

Ejemplo La función $h(x) = \sqrt{x^2+1}$ es derivable en $\mathbb{R}$, pues es la composición del polinomio $f(x) = x^2 + 1$ con la función $g(x) = \sqrt{x}$, y ambas funciones son derivables en sus dominios. Sabemos que $f'(x) = 2x$ y $g'(x) = 1/(2\sqrt{x})$. La regla de la cadena nos da que

$$h'(x) = g'(x^2+1)f'(x) = \frac{2x}{2\sqrt{x^2+1}} = \frac{x}{\sqrt{x^2+1}}.$$

■

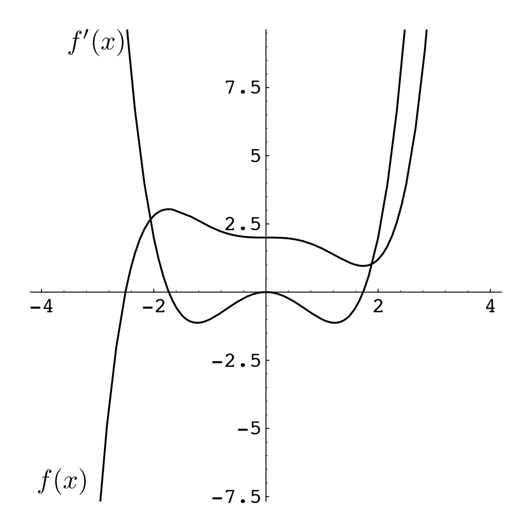
El lector que no esté familiarizado con el cálculo de derivadas debería practicar hasta que la derivación le resultara un acto mecánico. Hay muchos libros adecuados para ello, por lo que en adelante dejaremos de justificar los cálculos de derivadas.

3.3 Propiedades de las funciones derivables

La derivada de una función contiene mucha información sobre ésta. Consideremos por ejemplo

$$f(x) = \frac{x^5}{10} - \frac{x^3}{2} + 2, \quad f'(x) = \frac{x^4}{2} - \frac{3}{2}\,x^2.$$

La figura de la página siguiente muestra las gráficas. Fijémonos en el signo de la derivada. Es fácil ver que f' tiene una raíz doble en 0 y dos raíces simples en $\pm\sqrt{3}$. La restricción de f' a cada uno de los intervalos $\big]-\infty, -\sqrt{3}\big[$, $\big]-\sqrt{3}, 0\big[$, $\big]0, \sqrt{3}\big[$ y $\big]\sqrt{3}, +\infty\big[$ es una función continua que no se anula, luego por el teorema de los valores intermedios f' tiene signo constante en cada uno de ellos. Es fácil ver entonces que el signo de f' varía como indica la gráfica, es decir, f' es positiva en los dos intervalos no acotados y negativa en los acotados.



En los puntos donde f' es positiva la tangente a f tiene pendiente positiva, y vemos en la gráfica que esto se traduce en que f es creciente. Por el contrario, en los intervalos donde f' es negativa la función f es decreciente.

En los puntos donde f' se anula la tangente a f es horizontal. En $-\sqrt{3}$, la derivada pasa de ser positiva a ser negativa, luego f pasa de ser creciente a ser decreciente, y por ello el punto es un máximo relativo, en el sentido de que f toma en $-\sqrt{3}$ un valor mayor que en los puntos de alrededor. En cambio, en $\sqrt{3}$ la derivada pasa de negativa a positiva, f pasa de decreciente a creciente y el punto es un mínimo relativo. El caso del 0 es distinto, pues f' es negativa a la izquierda, toca el 0 y vuelve a bajar, con lo que sigue siendo negativa. Por ello f es creciente en 0 y no tiene ni un máximo ni un mínimo en 0.

Vemos así que conociendo la derivada podemos formarnos una idea de la función: dónde crece, dónde decrece, dónde tiene máximos y mínimos, y más cosas de las que no hemos hablado. Vamos a desarrollar todas estas ideas.

Definición 3.8 Sea $f : A \subset \mathbb{R} \longrightarrow \mathbb{R}$. Diremos que f tiene un *máximo relativo* en un punto $a \in A$ si existe un entorno V de a contenido en A de modo que para todo $x \in V$ se cumple $f(x) \leq f(a)$. Diremos que f tiene un *mínimo relativo* en a si existe un entorno V de a contenido en A tal que para todo $x \in V$ se cumple $f(a) \leq f(x)$. La función f tiene un *extremo relativo* en a si tiene un máximo o un mínimo relativo en a.

Teorema 3.9 *Si $f : A \longrightarrow \mathbb{R}$ es una función derivable en un punto $a \in A$ y f tiene un extremo relativo en a, entonces $f'(a) = 0$.*

Demostración: Supongamos, por reducción al absurdo, que $f'(a) > 0$. (El caso $f'(a) < 0$ se razona análogamente.) Entonces $]0, +\infty[$ es un entorno de $f'(a)$, luego por definición de límite y de derivada existe un $\epsilon > 0$ de manera que $]a - \epsilon, a + \epsilon[\subset A$ y si $0 < |h| < \epsilon$ entonces

$$\frac{f(a+h) - f(a)}{h} > 0.$$

Esto se traduce en que $f(a+h) > f(a)$ si $h > 0$ y $f(a+h) < f(a)$ si $h < 0$, lo que contradice que f tenga un extremo relativo en a. ■

La función del ejemplo anterior muestra que $f'(a) = 0$ no implica que a sea un extremo relativo. Más adelante volveremos sobre este punto. Ahora probemos una consecuencia sencilla de este teorema:

Teorema 3.10 (Teorema de Rolle) *Sea $f : [a, b] \longrightarrow \mathbb{R}$ una función continua en $[a, b]$ y derivable en $]a, b[$. Si $f(a) = f(b)$, entonces existe un $c \in\ ]a, b[$ tal que $f'(c) = 0$.*

DEMOSTRACIÓN: Como $[a, b]$ es compacto, la función f alcanza un valor mínimo m y un valor máximo M. Si se cumpliera que $m = M = f(a) = f(b)$, entonces f sería constante y su derivada sería nula, luego cualquier $c \in\ ]a, b[$ cumpliría el teorema.

Supongamos que $m < M$. Entonces, o bien $m \neq f(a)$ o bien $M \neq f(a)$. Digamos por ejemplo $M \neq f(a)$. Sea $c \in [a, b]$ el punto donde $f(c) = M$. Como $M \neq f(a) = f(b)$, ha de ser $a < b < c$, y como f toma en c su valor máximo, en particular c es un máximo relativo de f, luego por el teorema anterior $f'(c) = 0$.

■

Como aplicación podemos relacionar la derivabilidad y el crecimiento global de una función.

Teorema 3.11 *Sea A un intervalo abierto y $f : A \longrightarrow \mathbb{R}$ una función derivable en A tal que f' no se anule. Entonces f es estrictamente monótona en A y además se da uno de estos dos casos:*

a) o bien $f'(x) > 0$ para todo $x \in A$, y entonces f es estrictamente creciente en A,

b) o bien $f'(x) < 0$ para todo $x \in A$, y entonces f es monótona decreciente en A.

DEMOSTRACIÓN: En primer lugar, f es inyectiva, pues si $a < b$ son dos puntos de A tales que $f(a) = f(b)$, entonces existe un $c \in\ ]a, b[\ \subset A$ tal que $f'(c) = 0$, en contra de la hipótesis.

Por el teorema 3.5, f es estrictamente monótona en A. Si es monótona decreciente, la prueba del teorema 3.9 muestra que no puede ser $f'(a) > 0$ en ningún punto $a \in A$, luego ha de ser $f'(a) < 0$ en todos los puntos. Análogamente, si f es monótona creciente ha de ser $f'(a) > 0$ en toco $a \in A$. ■

Veamos ahora un resultado técnico:

Teorema 3.12 (Teorema de Cauchy) *Sean f, $g : [a, b] \longrightarrow \mathbb{R}$ funciones continuas en $[a, b]$ y derivables en $]a, b[$. Entonces existe un $c \in\ ]a, b[$ tal que*

$$g'(c)\big(f(b) - f(a)\big) = f'(c)\big(g(b) - g(a)\big).$$

DEMOSTRACIÓN: Consideremos la función dada por

$$h(x) = f(x)\big(g(b) - g(a)\big) - g(x)\big(f(b) - f(a)\big).$$

Se cumple que $h(a) = h(b) = f(a)g(b) - g(a)f(b)$. Además h es continua en $[a, b]$ y derivable en $]a, b[$. Por el teorema de Rolle existe un punto $c \in\,]a, b[$ tal que $h'(c) = 0$, pero $h'(x) = f'(x)\big(g(b) - g(a)\big) - g'(x)\big(f(b) - f(a)\big)$, luego

$$f'(c)\big(g(b) - g(a)\big) - g'(c)\big(f(b) - f(a)\big) = 0$$

■

Más adelante tendremos ocasión de usar este resultado en toda su generalidad, pero de momento nos basta con el caso particular que resulta de tomar como función g la dada por $g(x) = x$. Entonces tenemos:

Teorema 3.13 (Teorema del valor medio) *Sea $f : [a, b] \longrightarrow \mathbb{R}$ una función continua en $[a, b]$ y derivable en $]a, b[$. Entonces existe un $c \in\,]a, b[$ tal que*

$$f(b) - f(a) = f'(c)(b - a).$$

Notar que el teorema de Rolle es un caso particular del teorema del valor medio. Este teorema tiene una interpretación geométrica. La expresión

$$\frac{f(b) - f(a)}{b - a}$$

puede interpretarse como la "pendiente media" de f en $[a, b]$, es decir, es el cociente de lo que aumenta f cuando la variable x pasa de a a b dividido entre lo que ha aumentado la variable. Lo que dice el teorema del valor medio es que hay un punto en el intervalo donde la función toma el valor medio de su pendiente.

La importancia de este teorema es que nos relaciona una magnitud global, la pendiente media, con una magnitud local, la derivada en un punto. Las consecuencias son muchas. Una aplicación típica es el siguiente refinamiento del teorema 3.11:

Teorema 3.14 *Si $f : [a, b] \longrightarrow \mathbb{R}$ es continua en $[a, b]$, derivable en $]a, b[$ y su derivada es positiva (negativa) en $]a, b[$, entonces f es estrictamente creciente (decreciente) en $[a, b]$.*

DEMOSTRACIÓN: El teorema 3.11 nos da que f es estrictamente monótona en $]a, b[$. Sólo falta probar que es creciente o decreciente en a y en b.

Si $x \in\,]a, b[$, entonces $f(x) - f(a) = f'(c)(x - a)$, para cierto punto $c \in\,]a, x[$. Por lo tanto, si f' es positiva, $f(x) - f(a) > 0$ para todo $x \in\,]a, b[$, e igualmente se prueba que $f(b) - f(x) > 0$ para todo $x \in\,]a, b[$, luego f es creciente en $[a, b]$.

■

Sabemos que las funciones constantes tienen derivada nula. El teorema del valor medio nos da el recíproco:

Teorema 3.15 *Si una función tiene derivada nula en todos los puntos de un intervalo abierto, entonces es constante.*

DEMOSTRACIÓN: Sea f una función derivable en un intervalo A con derivada nula. Sean $a < b$ dos puntos cualesquiera de A. Entonces $f(b) - f(a) = f'(c)(b-a) = 0$, donde c es un punto de $]a, b[$. Por lo tanto f es constante. ■

Una consecuencia inmediata es el teorema siguiente, que afirma que una función derivable está unívocamente determinada por su derivada y su valor en un punto cualquiera.

Teorema 3.16 *Si f y g son funciones derivables en un intervalo abierto y $f' = g'$, entonces existe un $k \in \mathbb{R}$ tal que $f = g + k$.*

DEMOSTRACIÓN: La función $f - g$ tiene derivada nula, luego $f - g = k$. ■

Las derivadas proporcionan un teorema muy útil para el cálculo de límites. De momento no podemos estimar su valor porque los límites de las funciones que conocemos (polinomios, fracciones algebraicas. etc.) son fáciles de calcular directamente, pero más adelante tendremos ocasión de aprovecharlo.

Teorema 3.17 (Regla de L'Hôpital) *Sean f, $g :]a, b[\longrightarrow \mathbb{R}$ funciones derivables tales que $\lim_{x \to a} f(x) = \lim_{x \to a} g(x) = 0$ y de modo que g y g' no se anulen en $]a, b[$. Si existe*

$$\lim_{x \to a} \frac{f'(x)}{g'(x)} = L,$$

entonces también existe

$$\lim_{x \to a} \frac{f(x)}{g(x)} = L.$$

DEMOSTRACIÓN: Extendamos f y g al intervalo $[a, b[$ estableciendo que $f(a) = g(a) = 0$. Así siguen siendo continuas.

Si $a < x < b$, por el teorema de Cauchy existe un punto $c \in]a, x[$ tal que

$$\big(f(x) - f(a)\big)g'(c) = \big(g(x) - g(a)\big)f'(c),$$

o sea, $f(x)g'(c) = g(x)f'(c)$, y como $g(x) \neq 0 \neq g'(c)$, podemos escribir

$$\frac{f(x)}{g(x)} = \frac{f'(c)}{g'(c)}. \tag{3.2}$$

Por definición de límite, si $\epsilon > 0$, existe un $\delta > 0$ tal que si $0 < c - a < \delta$, entonces

$$\left| \frac{f'(c)}{g'(c)} - L \right| < \epsilon. \tag{3.3}$$

Así tenemos que si $0 < x - a < \delta$, existe un $c \in]a, x[$ que cumple (3.2) y (3.3). Por consiguiente, para todo $x \in]a, a + \delta[$ se cumple

$$\left| \frac{f(x)}{g(x)} - L \right| < \epsilon.$$

Esto significa que

$$\lim_{x \to a} \frac{f(x)}{g(x)} = L.$$

■

Obviamente la regla de L'Hôpital también es válida cuando en las hipótesis cambiamos a por b. Combinando las dos versiones obtenemos la regla de L'Hôpital para funciones definidas en intervalos $]a - \epsilon, a + \epsilon[\setminus \{a\}$ y tomando límites en a (si existe el límite del cociente de derivadas, existen los límites por la derecha y por la izquierda y coinciden, por los casos correspondientes de la regla, existen los límites de los cocientes de las funciones por ambos lados y coinciden, luego existe el límite y coincide con el de las derivadas).

Los teoremas siguientes demuestran otras variantes de la regla de L'Hôpital de no menor interés.

Teorema 3.18 (Regla de L'Hôpital) *Sean $f, g :]a, +\infty[\longrightarrow \mathbb{R}$ dos funciones derivables tales que $\lim_{x \to +\infty} f(x) = \lim_{x \to +\infty} g(x) = 0$ y de modo que g y g' no se anulan en $]a, +\infty[$. Si existe*

$$\lim_{x \to +\infty} \frac{f'(x)}{g'(x)} = L,$$

entonces también existe

$$\lim_{x \to +\infty} \frac{f(x)}{g(x)} = L.$$

DEMOSTRACIÓN: Consideremos las funciones $F(x) = f(1/x)$ y $G(x) = g(1/x)$, definidas en $]0, 1/a[$. Claramente F y G son continuas, y $\lim_{x \to 0} F(x) = \lim_{x \to 0} G(x) = 0$. Además por la regla de la cadena son funciones derivables y sus derivadas son

$$F'(x) = -\frac{f'(1/x)}{x^2}, \quad G'(x) = -\frac{g'(1/x)}{x^2}.$$

También es claro que ni G ni G' se anulan en su dominio y

$$\frac{F'(x)}{G'(x)} = \frac{f'(1/x)}{g'(1/x)},$$

luego existe

$$\lim_{x \to 0} \frac{F'(x)}{G'(x)} = L.$$

El caso ya probado de la regla de L'Hôpital nos da ahora que también existe

$$\lim_{x\to 0}\frac{F(x)}{G(x)} = L.$$

Obviamente entonces

$$\lim_{x\to +\infty}\frac{f(x)}{g(x)} = L.$$

■

Igualmente se prueba la regla de L'Hôpital para funciones definidas en intervalos $]-\ nfty, b[$ y cuando x tiende a $-\infty$.

Así pues, si tenemos una indeterminación de tipo $0/0$ y al derivar numerador y denominador podemos calcular el límite, la función original tiene ese mismo límite. Ahora veremos que la regla de L'Hôpital es aplicable también a indeterminaciones del tipo ∞/∞.

Teorema 3.19 (Regla de L'Hôpital) *Sean f, $g :]a, +\infty[\longrightarrow \mathbb{R}$ dos funciones derivables tales que $\lim\limits_{x\to+\infty} f(x) = \lim\limits_{x\to+\infty} g(x) = \infty$ y de modo que g y g' no se anulan en $]a, +\infty[$. Si existe*

$$\lim_{x\to +\infty}\frac{f'(x)}{g'(x)} = L,$$

entonces también existe

$$\lim_{x\to +\infty}\frac{f(x)}{g(x)} = L.$$

Demostración: Por definición de límite, dado $\epsilon > 0$, existe un $M > a$ tal que si $x > M$ entonces

$$\left|\frac{f'(x)}{g'(x)} - L\right| < \epsilon.$$

Por el teorema de Cauchy, si $x > M$, existe un $y \in]M, x[$ de modo que

$$\frac{f(x) - f(M)}{g(x) - g(M)} = \frac{f'(y)}{g'(y)},$$

luego

$$\left|\frac{f(x) - f(M)}{g(x) - g(M)} - L\right| < \epsilon.$$

(Notar que, como g' no se anula, la función g es monótona, luego el denominador es no nulo).

Puesto que $\lim\limits_{x\to+\infty} f(x) = \infty$, existe un $N > M$ tal que si $x > N$ entonces $|f(x)| > |f(M)|$, y en particular $f(x) - f(M) \neq 0$. Por ello, para $x > N$ podemos escribir

$$\frac{f(x)}{g(x)} = \frac{f(x) - f(M)}{g(x) - g(M)}\,\frac{f(x)}{f(x) - f(M)}\,\frac{g(x) - g(M)}{g(x)}.$$

Si en los dos últimos factores dividimos numerador y denominador entre $f(x)$ y $g(x)$ respectivamente, queda claro que tienden a 1 cuando $x \to +\infty$, luego tomando N suficientemente grande podemos suponer que si $x > N$ entonces el producto de ambos dista de 1 menos de ϵ. De este modo, para x suficientemente grande, el cociente $f(x)/g(x)$ se puede expresar como producto de dos números reales, uno arbitrariamente próximo a L y otro arbitrariamente próximo a 1. De la continuidad del producto se sigue que

$$\lim_{x\to+\infty} \frac{f(x)}{g(x)} = L.$$

■

Naturalmente la regla de L'Hôpital también es válida en el caso ∞/∞ cuando $x \to -\infty$. El mismo argumento que nos ha permitido pasar del caso finito al caso infinito en la indeterminación $0/0$ nos permite pasar ahora al caso finito. Es fácil probar:

Teorema 3.20 (Regla de L'Hôpital) *Sean f, $g :]a, b[\longrightarrow \mathbb{R}$ derivables tales que $\lim\limits_{x\to a} f(x) = \lim\limits_{x\to a} g(x) = \infty$ y de modo que g y g' no se anulan en $]a, b[$. Si existe*

$$\lim_{x\to a} \frac{f'(x)}{g'(x)} = L,$$

entonces también existe

$$\lim_{x\to a} \frac{f(x)}{g(x)} = L.$$

También se cumple la versión correspondiente cuando x tiende a b y cuando x tiende a un punto por la izquierda y la derecha a la vez.

3.4 La diferencial de una función

Consideremos una función $f : A \longrightarrow \mathbb{R}$ derivable en un punto $a \in A$. Sea Δx un número próximo a 0 de modo que $a + \Delta x \in A$ (el término Δx se lee "incremento de x", porque representa un pequeño aumento de la variable x en el punto a). Al calcular f en el punto $a + \Delta x$ obtenemos una variación o incremento de f dado por $\Delta_a(f) = f(a + \Delta x) - f(a)$. La expresión $\Delta_a(f)$ representa a una función de la variable Δx, definida en un entorno de 0. La derivada de f en a es, por definición,

$$f'(a) = \lim_{\Delta x\to 0} \frac{\Delta_a(f)}{\Delta x}.$$

Llamemos

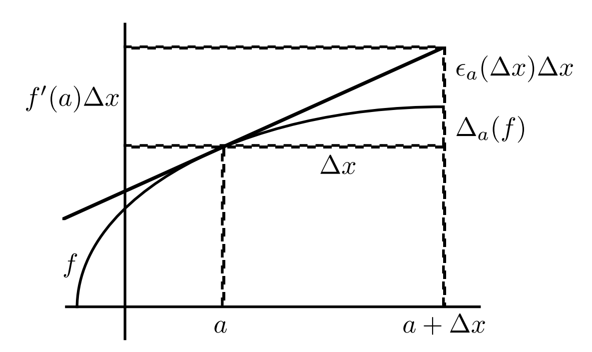
$$\epsilon_a(\Delta x) = \frac{\Delta_a(f)}{\Delta x} - f'(a).$$

Así, $\lim\limits_{\Delta x\to 0} \epsilon_a(\Delta x) = 0$.

Desarrollando las definiciones tenemos que

$$\Delta_a(f) = f(a + \Delta x) - f(a) = f'(a)\Delta x + \epsilon(\Delta x)\Delta x.$$

La figura muestra la situación:



El hecho de que $\epsilon_a(\Delta x)$ tienda a 0 expresa simplemente el hecho de que, para valores pequeños de Δx, se cumple $f(a+\Delta x) - f(a) \approx f'(a)\Delta x$, donde el signo $\approx$ significa "aproximadamente igual", es decir, que el valor de la función $f(a+\Delta x)$ es similar al de la recta tangente $f(a)+f'(a)\Delta x$. En la figura ambos valores son muy diferentes porque hemos tomado un Δx grande para mayor claridad.

Llamaremos *diferencial* de f en el punto a a la aplicación $df(a) : \mathbb{R} \longrightarrow \mathbb{R}$ dada por $df(a)(\Delta x) = f'(a)\Delta x$. Así $df(a)$ es una aplicación lineal en $\mathbb{R}$ de modo que $f(a + \Delta x) - f(a) \approx df(a)(\Delta x)$, o sea, la función $df(a)$ aproxima las diferencias entre las imágenes de f en puntos cercanos al punto a y la imagen de a. De aquí su nombre.

Si consideramos la función polinómica x, su derivada es 1, luego la diferencial de x es simplemente $dx(a)(\Delta x) = \Delta x$. Por ello podemos escribir $df(a)(\Delta x) = f'(a)\,dx(a)(\Delta x)$, luego tenemos la igualdad funcional

$$df(a) = f'(a)\,dx(a).$$

Si la función f es derivable en todo punto de A, la igualdad anterior se cumple en todo punto, luego si consideramos a df y dx como aplicaciones de A en el espacio de aplicaciones lineales de $\mathbb{R}$ en $\mathbb{R}$, tenemos la igualdad funcional

$$df = f'\,dx,$$

donde dx es la función constante que a cada $a \in A$ le asigna la función identidad en $\mathbb{R}$.

Del mismo modo que la estructura topológica permite hablar de los puntos de alrededor de un punto dado, pese a que ningún punto en particular está alrededor de otro, así mismo la diferencial de una función recoge el concepto de "incremento infinitesimal" de una función, pese a que ningún incremento en particular es infinitamente pequeño. Por ejemplo, la igualdad $dx^2 = 2x\,dx$

expresa que cuando la variable x experimenta un incremento infinitesimal dx, la función x^2 experimenta un incremento infinitesimal de $2x\,dx$. Con rigor, dx^2 no es un incremento infinitesimal, sino la función que a cada incremento Δx le asigna una aproximación al incremento correspondiente de x^2, de modo que lo que propiamente tenemos es la aproximación que resulta de evaluar dx en incrementos concretos, es decir, $\Delta_x(x^2) \approx 2x\Delta x$. El error de esta aproximación se puede hacer arbitrariamente pequeño tomando Δx suficientemente pequeño.

Por ejemplo, $(1{,}1)^2 \approx 1^2 + dx^2(1)(0{,}1) = 1 + 2 \cdot 0,1 = 1{,}2$. En realidad $(1{,}1)^2 = 1{,}21$, luego el error cometido es de una centésima.

Dada la igualdad $df = f'\,dx$, representaremos también la derivada de f mediante la notación

$$f'(x) = \frac{df}{dx},$$

que expresa que $f'(x)$ es la proporción entre las funciones df y dx, o también que $f'(x)$ es la razón entre un incremento infinitesimal de f respecto al incremento infinitesimal de x que lo ocasiona.

Es costumbre, especialmente en física, nombrar las funciones, no por la expresión que las determina, sino por la magnitud que determinan. Por ejemplo, supongamos que la posición e de un objeto depende del tiempo viene dada por la relación $e(t) = t^2$ Entonces la velocidad del móvil es

$$v(t) = \frac{de}{dt} = 2t,$$

lo que nos permite expresar t en función de v, mediante $t(v) = v/2$. A su vez, esto nos permite calcular la posición en función de la velocidad, mediante la función $e(v) = v^2/4$.

De este modo, llamamos e tanto a la función $e(t)$ como a la función $e(v)$, que son funciones distintas. La letra v representa a una función en $v = 2t$ y a una variable en $t = v/2$. Estos convenios no provocan ninguna ambigüedad, al contrario, en muchos casos resultan más claros y permiten expresar los resultados de forma más elegante. Por ejemplo, si tenemos dos funciones $y = y(x)$ y $z = z(y)$, entonces la función compuesta se expresa, en estos términos, como $z = z(x)$. Si las funciones son derivables en sus dominios, la regla de la cadena se convierte en

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}.$$

La primera derivada es la de la función compuesta $z(x)$, mientras que la segunda es la de $z(y)$. No es necesario indicar que dicha derivada ha de calcularse en $y(x)$, pues esto ya está implícito en el hecho de que se trata de una función de y (no de x). Por ejemplo,

$$\frac{de}{dt} = \frac{de}{dv}\frac{dv}{dt} = \frac{v}{2} \cdot 2 = v = 2t,$$

como cabía esperar.

Similarmente, si $y(x)$ es una función inyectiva y derivable con derivada no nula, su inversa se representa por $x(y)$, y el teorema de la función inversa se expresa así:

$$\frac{dy}{dx} = \frac{1}{\frac{dx}{dy}}.$$

Por ejemplo,

$$\frac{dv}{dt} = \frac{1}{\frac{dt}{dv}} = \frac{1}{1/2} = 2.$$

Vemos, pues, que en estos términos las propiedades de las derivadas son formalmente análogas a las de las fracciones.

3.5 El teorema de Taylor

Sea $f : A \longrightarrow \mathbb{R}$ una función derivable en el abierto A y tal que $f' : A \longrightarrow \mathbb{R}$ también sea derivable en A. Entonces a la derivada de f' se la denomina *derivada segunda* de f en A, y se representa por f''.

A su vez la derivada segunda puede ser derivable, y entonces está definida la derivada tercera, y así sucesivamente. Si una función admite n derivadas en A, a la derivada n-sima se la representa por $f^{n)} : A \longrightarrow \mathbb{R}$.

Conviene usar la notación $f^{0)}$ para referirse a la propia función f.

Llamaremos $C^n(A)$ al conjunto de las funciones definidas en A que admiten n derivadas y todas ellas son continuas en A. Si llamamos $C^0(A) = C(A)$, es decir, al conjunto de las funciones continuas en A, entonces tenemos

$$C^0(A) \supset C^1(A) \supset C^2(A) \supset C^3(A) \supset C^4(A) \supset \cdots$$

Llamaremos $C^\infty(A)$ al conjunto de las funciones infinitamente derivables en A. Por ejemplo, los polinomios y las fracciones algebraicas son de clase C^∞ en su dominio. Es inmediato que estos conjuntos son todos subálgebras de $C(A)$.

Las inclusiones son todas estrictas. Por ejemplo, si $a \in A$ es fácil ver que la función dada por

$$f(x) = \begin{cases} (x-a)^{n+1} & \text{si } x \geq a \\ -(x-a)^{n+1} & \text{si } x \leq a \end{cases}$$

es una función de clase $C^n(A)$ pero no de clase $C^{n+1}(A)$.

Según sabemos, si una función f es derivable en un punto a, entonces alrededor de a la función f puede ser aproximada por su recta tangente, esto es, por el polinomio $f(a) + f'(a)(x-a)$. La recta tangente es el único polinomio $P(x)$ de grado 1 que cumple $P(a) = f(a)$ y $P'(a) = f'(a)$.

Cabe suponer que si una función f admite dos derivadas y tomamos un polinomio P de grado 2 tal que $P(a) = f(a)$, $P'(a) = f'(a)$ y $P''(a) = f''(a)$, el polinomio P nos dará una aproximación mejor de la función f que la recta tangente. Esto no siempre es así, pero hay bastante de verdad en ello. Vamos a investigarlo.

Ante todo, si K es un cuerpo y $a \in K$, la aplicación $u : K[x] \longrightarrow K[x]$ dada por $u(p) = p(x-a)$ es un isomorfismo de K-espacios vectoriales. Como los polinomios $1, x, x^2, x^3, \ldots$ son una K-base de $K[x]$, resulta que los polinomios

$$1, \quad (x-a), \quad (x-a)^2, \quad (x-a)^3, \quad (x-a)^4, \ldots$$

son también una K-base, es decir, que todo polinomio de $K[x]$ se expresa de forma única como

$$P(x) = c_0 + c_1(x-a) + c_2(x-a)^2 + \cdots + c_n(x-a)^n, \tag{3.4}$$

para cierto natural n y ciertos coeficientes $c_o, \ldots, c_n \in K$.

Si una función f admite n derivadas en un punto a, las ecuaciones

$$f(a) = P(a), \quad f'(a) = P'(a), \quad \ldots \quad f^{n)}(a) = P^{n)}(a)$$

son satisfechas por un único polinomio de grado $\leq n$. En efecto, si $P(x)$ viene dado por (3.4), entonces $P(a) = c_0$, luego ha de ser $c_0 = f(a)$. Derivando obtenemos

$$P'(x) = c_1 + 2c_2(x-a) + \cdots + nc_n(x-a)^{n-1},$$

de donde $P'(a) = c_1$, y ha de ser $c_1 = f'(a)$. Similarmente, $P''(a) = 2c_2$, luego $c_2 = f''(a)/2$. Igualmente se obtiene $c_3 = f'''(a)/6$ y, en general, $c_k = f^{k)}(a)/k!$. En resumen:

$$P(x) = \sum_{k=0}^{n} \frac{f^{k)}}{k!}(x-a)^k.$$

Recíprocamente, es fácil ver que el polinomio $P(x)$ así definido cumple que $P^{k)}(a) = f^{k)}(a)$ para $k = 0, \ldots, n$.

Definición 3.21 Sea f una función derivable n veces en un punto a. Llamaremos *polinomio de Taylor* de grado n de f en a al polinomio

$$P_n(f)(x) = \sum_{k=0}^{n} \frac{f^{k)}}{k!}(x-a)^k \in \mathbb{R}[x].$$

El polinomio de Taylor es el único polinomio P de grado menor o igual que n que cumple $P^{k)}(a) = f^{k)}(a)$ para $k = 0, \ldots, n$. En particular si f es un polinomio de grado menor o igual que n se ha de cumplir $P_n(f) = f$.

Notar también que $P_0(f) = f(a)$, y que $P_1(f) = f(a) + f'(a)(x-a)$ es la recta tangente a f en a. Nuestra conjetura es que $P_n(f)$ es el polinomio de grado menor o igual que n que más se parece a f alrededor de a.

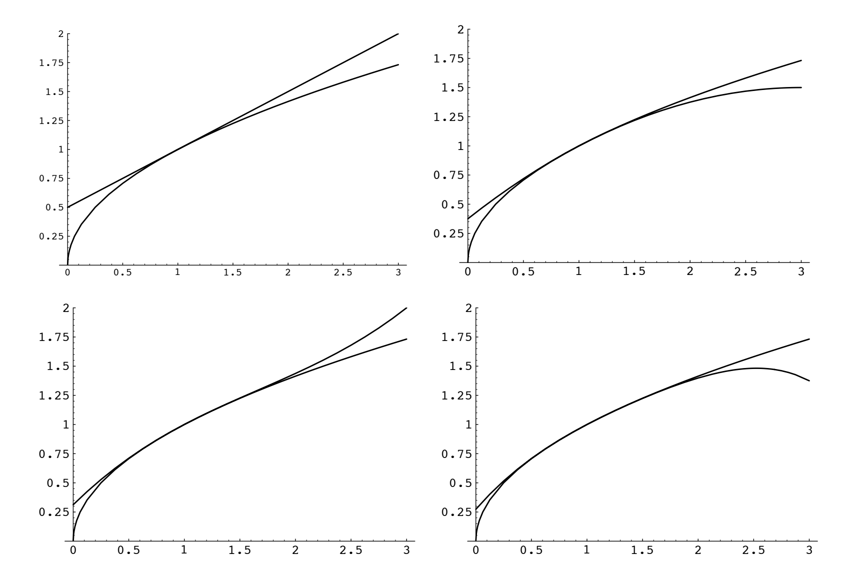
Ejemplo Consideremos la función $f(x) = x^{1/2}$ y $a = 1$. Calculemos sus derivadas:

Orden	Derivada	en 1
0	$x^{1/2}$	1
1	$\frac{1}{2}x^{-1/2}$	$\frac{1}{2}$
2	$-\frac{1}{2}\,\frac{1}{2}x^{-3/2}$	$-\frac{1}{2}\,\frac{1}{2}$
3	$\frac{1}{2}\,\frac{1}{2}\,\frac{3}{2}x^{-5/2}$	$\frac{1}{2}\,\frac{1}{2}\,\frac{3}{2}$
4	$-\frac{1}{2}\,\frac{1}{2}\,\frac{3}{2}\,\frac{5}{2}x^{-7/2}$	$-\frac{1}{2}\,\frac{1}{2}\,\frac{3}{2}\,\frac{5}{2}$

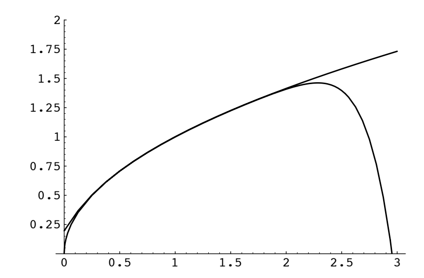
En general se prueba que las derivadas en 1 van alternando el signo, en el numerador tienen el producto de los primeros impares y en el denominador las sucesivas potencias de 2. Con esto podemos calcular cualquier polinomio de Taylor en 1:

$$
\begin{aligned}
P_0(f)(x) &= 1, \\
P_1(f)(x) &= 1 + \frac{1}{2}(x-1), \\
P_2(f)(x) &= 1 + \frac{1}{2}(x-1) - \frac{1}{8}(x-1)^2, \\
P_3(f)(x) &= 1 + \frac{1}{2}(x-1) - \frac{1}{8}(x-1)^2 + \frac{1}{16}(x-1)^3.
\end{aligned}
$$

Aquí están sus gráficas junto a la de la función. Vemos que el intervalo en que se confunden con la gráfica de f es cada vez mayor.



El polinomio de grado 8 es bastante representativo de lo que sucede cuando n es grande:



Vemos que la aproximación es cada vez mejor en el intervalo $[0, 2]$, pero a partir del 2 el polinomio se aleja bruscamente. Un ejemplo numérico:

$$P_8(f)(1,5) = 1,224729895, \quad \text{mientras que } \sqrt{1,5} = 1,224744871\ldots$$

La aproximación tiene 5 cifras exactas. ■

Los polinomios de Taylor plantean varios problemas importantes. La cuestión principal es si el error producido al aproximar una función de clase C^∞ por sus polinomios de Taylor puede reducirse arbitrariamente aumentando suficientemente el grado.

Definición 3.22 Si f es una función de clase C^∞ en un entorno de un punto a, llamaremos *serie de Taylor* de f en a a la serie funcional

$$\sum_{k=0}^{\infty} \frac{f^{k)}(a)}{k!}(x-a)^k.$$

La cuestión es si la serie de Taylor de f converge a f. El ejemplo anterior sugiere que la serie de $\sqrt{x}$ en 1 converge a la función en el intervalo $]0, 2[$, pero no parece converger más allá de 2. También hemos de tener presente la posibilidad de que la serie de Taylor de una función f converja a una función distinta de f. Para estudiar estos problemas introducimos el concepto de resto de Taylor:

Sea f una función derivable n veces en un punto a. Llamaremos *resto de Taylor* de grado n de f en a, a la función $R_n(f)(x) = f(x) - P_n(f)(x)$, donde $P_n(f)$ es el polinomio de Taylor de grado n de f en a.

Nuestro problema es determinar el comportamiento del resto de una función. Para ello contamos con el siguiente teorema, que es una generalización del teorema del valor medio.

Teorema 3.23 (Teorema de Taylor) *Sea $f : A \longrightarrow \mathbb{R}$ una función derivable $n+1$ veces en un intervalo abierto A y $a \in A$. Entonces para cada $x \in A$ existe un $\lambda \in\,]0, 1[$ tal que si $c = \lambda a + (1-\lambda)x$, se cumple*

$$R_n(f)(x) = \frac{f^{n+1)}(c)}{(n+1)!}(x-a)^{n+1}.$$

Demostración: Para $x = a$ es evidente, pues se cumple $P_n(f)(a) = f(a)$ y $R_n(f)(x) = 0$. Supongamos que $x \neq a$. Sea

$$Q(x) = \frac{1}{(x-a)^{n+1}} R_n(f)(x).$$

Sea $F : A \longrightarrow \mathbb{R}$ dada por

$$\begin{aligned} F(t) &= f(x) - \Big(f(t) + \frac{x-t}{1!} f'(t) + \frac{(x-t)^2}{2!} f''(t) + \cdots + \frac{(x-t)^n}{n!} f^{n)}(t) \\ &\quad + (x-t)^{n+1} Q(x) \Big). \end{aligned}$$

La función f y sus n primeras derivadas son continuas y derivables, luego F también es continua y derivable en A. Además $F(x) = f(x) - f(x) = 0$ y

$$F(a) = f(a) - \big(P_n(f)(a) + (x-a)^{n+1} Q(x) \big) = R_n(f)(x) - R_n(f)(x) = 0.$$

Por el teorema de Rolle existe un punto entre a y x, o sea, de la forma $c = \lambda a + (1-\lambda)x$, tal que $F'(c) = 0$. Calculemos en general $F'(t)$:

$$\begin{aligned} F'(t) &= 0 - \Big(f'(t) - f'(t) + \frac{x-t}{1!} f''(t) - \frac{2(x-t)}{2!} f''(t) + \frac{(x-t)^2}{2!} f'''(t) \\ &\quad \cdots - \frac{n(x-t)^{n-1}}{n!} f^{n)}(t) + \frac{(x-t)^n}{n!} f^{n+1)}(t) - (n+1)(x-t)^n Q(x) \Big). \end{aligned}$$

Los términos consecutivos se cancelan entre sí, y queda

$$F'(t) = -\frac{(x-t)^n}{n!} f^{n+1)}(t) + (n+1)(x-t)^n Q(x).$$

Como $F'(c) = 0$, evaluando en c queda

$$Q(x) = \frac{f^{n+1)}(c)}{(n+1)!},$$

y por definición de Q:

$$R_n(f)(x) = \frac{f^{n+1)}(c)}{(n+1)!} (x-a)^{n+1}.$$

■

Así pues, la diferencia entre $P_n(f)(x)$ y $f(x)$ tiene la forma de un monomio más del polinomio de Taylor salvo por el hecho de que la derivada $(n+1)$-ésima no se evalúa en el punto a, sino en un punto intermedio entre a y x.

Por ejemplo, si las derivadas de f están uniformemente acotadas en un intervalo A, es decir, si existe una misma constante K tal que $|f^{n)}(x)| \leq K$ para todo natural n y para todo $x \in A$, entonces

$$|f(x) - P_n(f)(x)| = \left| \frac{f^{n+1)}(c)}{(n+1)!} (x-a)^{n+1} \right| \leq \frac{K|x-a|^{n+1}}{(n+1)!}.$$

En el ejemplo de la página 56 probamos que la sucesión $M^n/n!$ converge a 0, luego la sucesión $\{P_n(f)(x)\}_{n=0}^{\infty}$ tiende a $f(x)$. Así pues:

Teorema 3.24 *Si A es un intervalo abierto, $a \in A$, $f \in C^\infty(A)$ y las derivadas de f están uniformemente acotadas en A, entonces para cada punto $x \in A$ se cumple*

$$f(x) = \sum_{n=0}^{\infty} \frac{f^{n)}(a)}{n!}(x-a)^n.$$

Este teorema no es aplicable a $\sqrt{x}$. En muchos casos, entre ellos el de esta función, los problemas de convergencia de las series de Taylor se vuelven evidentes en el contexto de la teoría de funciones de variable compleja (ver el capítulo XII). Los resultados de la sección siguiente resultan de gran ayuda en muchos casos, como tendremos ocasión de comprobar más adelante.

3.6 Series de potencias

Definición 3.25 Sea $a \in \mathbb{C}$ y $\{a_n\}_{n=0}^{\infty}$ una sucesión en $\mathbb{C}$. La *serie de potencias* de coeficientes $\{a_n\}_{n=0}^{\infty}$ y centro a es la serie funcional

$$\sum_{n=0}^{\infty} a_n(z-a)^n.$$

Las series de Taylor son, pues, series de potencias. En muchos casos es fácil determinar en qué puntos converge una serie de potencias. Para verlo necesitamos el concepto de límite superior de una sucesión de números reales. Se trata de lo siguiente:

Sea $\{a_n\}_{n=0}^{\infty}$ una sucesión de números reales. Su *límite superior* es el supremo (en $\overline{\mathbb{R}}$) del conjunto de sus puntos adherentes. Lo representaremos mediante $\overline{\lim_n}\, a_n$.

Se cumple que

$$\overline{\lim_n}\, a_n = \inf_{k\geq 0} \sup_{n\geq k} a_n.$$

En efecto, sea p un punto adherente de $\{a_n\}_{n=0}^{\infty}$. Dados $\epsilon > 0$ y $k \geq 0$, existe un $n \geq k$ tal que $a_n \in \,]p-\epsilon, p+\epsilon[$, luego $p - \epsilon \leq \sup_{n\geq k} a_n$. Esto vale para todo $\epsilon > 0$, luego $p \leq \sup_{n\geq k} a_n$ para todo $k \geq 0$, luego $p \leq \inf_{k\geq 0} \sup_{n\geq k} a_n$. Como el límite superior es el supremo de estos p, tenemos que $\overline{\lim_n}\, a_n \leq \inf_{k\geq 0} \sup_{n\geq k} a_n$.

Sea $L = \inf_{k\geq 0} \sup_{n\geq k} a_n$. Dado $\epsilon > 0$, existe un $k \geq 0$ tal que $L \leq \sup_{n\geq k} a_n \leq L+\epsilon$. Si $L = \sup_{n\geq k} a_n$, entonces existe un $n \geq k$ tal que $L - \epsilon < a_n \leq L$. Si por el contrario $L < \sup_{n\geq k} a_n < L + \epsilon$, entonces existe un $n \geq k$ tal que $L \leq a_n < L + \epsilon$.

En cualquier caso existe un $n \geq k$ tal que $a_n \in \,]L-\epsilon, L+\epsilon[$. Esto significa que L es un punto adherente de la sucesión, luego $L \leq \overline{\lim_n}\, a_n$ y tenemos la igualdad. ■

Por la propia definición es claro que si una sucesión converge en $\overline{\mathbb{R}}$ entonces su límite, que es su único punto adherente, coincide con su límite superior. Ahora podemos probar:

Teorema 3.26 *Sea* $\sum\limits_{n=0}^{\infty} a_n(z-a)^n$ *una serie de potencias, sea* $R = 1/\overline{\lim\limits_n}\sqrt[n]{|a_n|}$ *(entendiendo que* $1/0 = +\infty$ *y* $1/(+\infty) = 0$*). Entonces la serie converge absoluta y uniformemente en todo compacto contenido en* $B_R(a)$ *y diverge en todo punto de* $\mathbb{C} \setminus \overline{B}_R(a)$ *(las bolas se toman respecto a la norma euclídea. Convenimos que* $B_{+\infty}(a) = \mathbb{C}$*). En particular la serie converge absoluta y puntualmente en* $B_R(a)$ *a una función continua.*

Demostración: Sea K un compacto en $B_R(a)$. Veamos que la serie converge absoluta y uniformemente en K. La función $|x-a|$ es continua en K, luego alcanza su máximo r en un punto $x \in K$, es decir, $|x-a| = r$ y para todo $y \in K$ se cumple $|y-a| \leq r$. Así $K \subset \overline{B}_r(a)$.

Como $x \in B_R(a)$ ha de ser $r < R$, luego $r\,\overline{\lim\limits_n}\sqrt[n]{|a_n|} < 1$. Tomemos ρ tal que $r\,\overline{\lim\limits_n}\sqrt[n]{|a_n|} < \rho < 1$. Como $\overline{\lim\limits_n}\sqrt[n]{|a_n|} = \inf\limits_{k\geq 0}\sup\limits_{n\geq k}\sqrt[n]{|a_n|}$, existe un natural k tal que $\sup\limits_{n\geq k}\sqrt[n]{|a_n|} < \rho/r$, luego si $n \geq k$ se cumple $\sqrt[n]{|a_n|} < \rho/r$ y por lo tanto $|a_n|r^n < \rho^n$.

Si $y \in K$ entonces $|y-a| \leq r$, luego $|y-a|^n \leq r^n$, luego $|a_n(y-a)^n| \leq |a_n|r^n < \rho^n$. Así pues, la serie $\sum\limits_{n=k}^{\infty} a_n(y-a)^n$ está mayorada en K por $\sum\limits_{n=k}^{\infty} \rho^n$, que es convergente por ser geométrica de razón menor que 1. El criterio de Weierstrass nos da que la serie de potencias converge absoluta y uniformemente a una función continua en K. Todo punto de $B_R(a)$ tiene un entorno compacto contenido en $B_R(a)$ (una bola cerrada de radio adecuado), luego la suma es continua en $B_R(a)$.

Ahora veamos que la serie diverge en $\mathbb{C}\setminus\overline{B}_R(a)$. Sea $x \in \mathbb{C}$ tal que $|x-a| > R$. Entonces $1 < |x-a|\,\overline{\lim\limits_n}\sqrt[n]{|a_n|}$. Por lo tanto, para todo natural k se cumple $1/|x-a| < \sup\limits_{n\geq k}\sqrt[n]{|a_n|}$, luego existe un $n \geq k$ tal que $\sqrt[n]{|a_n|}|x-a| > 1$, o sea, $|a_n(y-a)^n| > 1$. Esto significa que $a_n(y-a)^n$ no tiende a 0, luego la serie diverge. ■

El número R se llama *radio de convergencia* de la serie de potencias. La bola $B_R(a)$ se llama *disco de convergencia.* Tenemos, pues que una serie de potencias converge absolutamente en su disco de convergencia y diverge en los puntos exteriores a él (los puntos interiores de su complementario). En cada punto de la frontera del disco la serie puede converger absolutamente, condicionalmente o diverger, según los casos.

A la hora de determinar el radio de convergencia de una serie suele ser útil el teorema siguiente:

Teorema 3.27 *Sea* $\sum\limits_{n=0}^{\infty} a_n(z-a)^n$ *una serie de potencias tal que exista*

$$\lim_n \frac{|a_{n+1}|}{|a_n|} = L.$$

Entonces su radio de convergencia es $1/L$.

DEMOSTRACIÓN: Por el teorema anterior, el radio de convergencia de la serie dada es el mismo que el de la serie $\sum\limits_{n=0}^{\infty} |a_n|z^n$. Si $x > 0$, tenemos que

$$\lim_n \frac{|a_{n+1}|x^{n+1}}{|a_n|x^n} = Lx,$$

luego el criterio de D'Alembert implica que la serie converge cuando $Lx < 1$ y diverge si $LX > 1$. Consecuentemente el radio de convergencia ha de ser $1/L$. ■

Las series de potencias se pueden derivar término a término. Conviene probar un resultado un poco más general:

Teorema 3.28 *Sea* $\{f_n\}_{n=1}^{\infty}$ *una sucesión de funciones* $f_n : \,]a,b[\, \longrightarrow \mathbb{R}$ *que converge uniformemente a una función* f. *Supongamos que todas ellas son derivables en* $]a,b[$ *y que la sucesión de derivadas converge uniformemente a una función* g. *Entonces* f *es derivable y* $g = f'$.

DEMOSTRACIÓN: Fijemos un punto $x \in \,]a,b[$ y consideremos las funciones

$$F_n(y) = \begin{cases} \frac{f_n(x)-f_n(y)}{x-y} & \text{si } y \neq x \\ f_n'(x) & \text{si } y = x \end{cases}$$

Similarmente definimos $F : \,]a,b[\, \longrightarrow \mathbb{R}$ mediante

$$F(y) = \begin{cases} \frac{f(x)-f(y)}{x-y} & \text{si } y \neq x \\ g(x) & \text{si } y = x \end{cases}$$

El hecho de que f_n sea derivable en x implica que F_n es continua en $]a,b[$. Basta probar que para todo $\epsilon > 0$ existe un número natural n_0 tal que si m, $n \geq n_0$ e $y \in \,]a,b[$ entonces $|F_m(y)-F_n(y)| < \epsilon$. En efecto, esto significa que $\{F_n\}_{n=1}^{\infty}$ es (uniformemente) de Cauchy, luego según el teorema 2.54 la sucesión ha de converger uniformemente a alguna función continua, pero dado que converge puntualmente a F, de hecho convergerá uniformemente a F. Tenemos, pues, que F es continua y a su vez esto implica que f es derivable en x y $f'(x) = g(x)$.

Existe un número natural n_0 tal que si m, $n > n_0$ entonces

$$|f_n'(u) - f_m'(u)| < \frac{\epsilon}{2} \quad \text{para todo } u \in \,]a,b[\,.$$

Entonces, dados $y \in]a, b[$ y m, $n \geq n_0$, si $y \neq x$ se cumple

$$\begin{aligned} |F_n(y) - F_m(y)| &= \left| \frac{f_n(x) - f_n(y)}{x - y} - \frac{f_m(x) - f_m(y)}{x - y} \right| \\ &\leq \left| \frac{1}{x - y} \right| |f_n(x) - f_m(x) - f_n(y) + f_m(y)| \end{aligned}$$

Aplicamos el teorema del valor medio a la función $f_n - f_m$ en el intervalo $[y, x]$ (suponemos, por ejemplo, $y < x$). Entonces existe un $u \in]a, b[$ tal que

$$f_n(x) - f_m(x) - f_n(y) + f_m(y) = \big(f'_n(u) - f'_m(u)\big)(x - y).$$

Así pues,

$$|F_n(y) - F_m(y)| \leq \left| \frac{1}{x - y} \right| |f'_n(u) - f'_m(u)| \, |x - y| < \frac{\epsilon}{2}.$$

Así mismo, si $y = x$ tenemos

$$|F_n(x) - F_m(x)| = |f'_n(x) - f'_m(x)| < \frac{\epsilon}{2} < \epsilon.$$

■

Como consecuencia tenemos:

Teorema 3.29 *Sea* $\sum\limits_{n=0}^{\infty} a_n(x - a)^n$ *una serie de potencias con centro y coeficientes reales. Supongamos que tanto ella como la serie* $\sum\limits_{n=1}^{\infty} na_n(x - a)^{n-1}$ *convergen en un intervalo* $]a - \epsilon, a + \epsilon[$*. Entonces la segunda serie es la derivada de la primera.*

Llamemos $f(x)$ a la función definida sobre $]a - \epsilon, a + \epsilon[$ por la serie dada y $g(x)$ a la función definida por la segunda serie. Hemos de probar que $f'(x) = g(x)$ para todo x en el intervalo.

Sea $f_n(x) = \sum\limits_{k=0}^{n} a_n(x - a)^n$. Se trata de una sucesión de polinomios cuyas derivadas $f'_n(x)$ son las sumas parciales de la segunda serie. Dado un punto $x \in]a - \epsilon, a + \epsilon[$, tomamos un intervalo cerrado $[x - \delta, x + \delta] \subset]a - \epsilon, a + \epsilon[$. Por el teorema 3.26, en este intervalo las sucesiones $f_n(x)$ y $f'_n(x)$ convergen absoluta y uniformemente a f y g respectivamente. ■

El lector puede demostrar que, en realidad, las dos series del teorema anterior tienen el mismo radio de convergencia, con lo que la hipótesis sobre la convergencia de la segunda es redundante. En la práctica no necesitaremos este hecho, pues es obvio que si existe $\lim\limits_{n} \frac{|a_{n+1}|}{|a_n|} = L$, entonces el límite correspondiente a la segunda serie también existe y vale lo mismo.

Con los resultados que veremos en la sección siguiente sobre la función exponencial es fácil probar esto mismo pero sobre $\overline{\lim\limits_{n}} \sqrt[n]{|a_n|}$, con lo que obtenemos

la igualdad de los radios de convergencia en el caso general. De aquí se sigue inmediatamente que una serie de potencias con coeficientes y centro reales es una función de clase C^∞ en su intervalo real de convergencia, y además coincide con su serie de Taylor.

3.7 La función exponencial

Vamos a aplicar las ideas de las secciones precedentes a la construcción de las funciones más importantes del análisis: la exponencial, la logarítmica y las trigonométricas. En esta sección nos ocuparemos de la función exponencial.

Hasta ahora tenemos definido a^r cuando a es un número real positivo y r es un número racional. No es difícil probar que la función $r \mapsto a^r$ admite una única extensión continua a $\mathbb{R}$ que sigue conservando la propiedad $a^{x+y} = a^x a^y$. Además esta función es infinitamente derivable y coincide en todo punto con su serie de Taylor en 0. En lugar de probar todos estos hechos lo que haremos será definir la función exponencial a partir de su serie de Taylor, para lo cual no necesitaremos siquiera el hecho de que ya la tenemos definida sobre $\mathbb{Q}$. No obstante, ahora vamos a suponer la existencia de la función a^x, así como que es derivable, y vamos a calcular su serie de Taylor. Así obtendremos la serie que deberemos tomar como definición.

Sea $f(x) = a^x$. Entonces,

$$f'(x) = \lim_{h\to 0} \frac{a^{x+h} - a^x}{h} = a^x \lim_{h\to 0} \frac{a^h - 1}{h} = a^x f'(0).$$

Llamemos $k = f'(0)$. Entonces hemos probado que $f'(x) = kf(x)$, luego por inducción concluimos que f es infinitamente derivable y $f^{n)}(x) = k^n f(x)$. No puede ser $k = 0$, o de lo contrario f sería constante. Sea $e = a^{1/k} = f(1/k)$. Entonces la función $g(x) = e^x = a^{x/k} = f(x/k)$ cumple $g'(x) = f'(x/k)(1/k) = f(x/k) = g(x)$, es decir, escogiendo adecuadamente la base e obtenemos una función exponencial que coincide con su derivada. Su serie de Taylor en 0 es entonces fácil de calcular: todas las derivadas valen $e^0 = 1$, lo cual nos lleva a la definición siguiente:

Definición 3.30 Llamaremos *función exponencial* a la definida por la serie de potencias

$$e^z = \sum_{n=0}^{\infty} \frac{z^n}{n!}.$$

Puesto que

$$\lim_n \frac{1/(n+1)!}{1/n!} = \lim_n \frac{1}{n} = 0,$$

el radio de convergencia es infinito, luego la exponencial está definida sobre todo número complejo z. El último teorema de la sección anterior implica que la exponencial real es derivable, y su derivada en un punto x es

$$\sum_{n=1}^{\infty} \frac{nx^{n-1}}{n!} = \sum_{n=1}^{\infty} \frac{x^{n-1}}{(n-1)!} = \sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x.$$

Es claro que $e^0 = 1$. Definimos el número

$$e = e^1 = \sum_{n=0}^{\infty} \frac{1}{n!} = 2,71828182845904523536028 74\ldots$$

Ahora probamos la ecuación que caracteriza a la función exponencial:

Teorema 3.31 *Si z_1, $z_2 \in \mathbb{C}$, entonces $e^{z_1+z_2} = e^{z_1}e^{z_2}$.*

DEMOSTRACIÓN: Usamos la fórmula del producto de Cauchy:

$$\begin{aligned} e^{z_1}e^{z_2} &= \sum_{n=0}^{\infty} \frac{z_1^n}{n!} \sum_{n=0}^{\infty} \frac{z_2^n}{n!} = \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{1}{k!\,(n-k)!} z_1^k z_2^{n-k} \\ &= \sum_{n=0}^{\infty} \sum_{k=0}^{n} \frac{1}{n!} \binom{n}{k} z_1^k z_2^{n-k} = \sum_{n=0}^{\infty} \frac{(z_1+z_2)^n}{n!} = e^{z_1+z_2}. \end{aligned}$$

■

De aquí obtenemos muchas consecuencias. Por una parte, si n es un número natural no nulo entonces $e^n = e^{1+\overset{n}{\cdots}+1} = e \overset{n}{\cdots} e$, es decir, la función exponencial sobre números naturales (incluido el 0) coincide con la exponenciación usual con base e. Así mismo, $1 = e^0 = e^{x-x} = e^x e^{-x}$, luego $e^{-x} = 1/e^x$, con lo que la función exponencial coincide también con la usual cuando el exponente es entero.

Como los coeficientes de la serie exponencial son positivos, vemos que si $x \geq 0$ entonces $e^x > 0$, y si $x < 0$ entonces $e^x = 1/e^{-x} > 0$. Así pues, $e^x > 0$ para todo número real x.

Como, $(e^{1/n})^n = e^{1/n+\cdots+1/n} = e^1 = e$, resulta que $e^{1/n} = \sqrt[n]{e}$. Es fácil ver ahora que $e^{p/q} = \sqrt[q]{e^p}$, luego la función exponencial coincide con la que teníamos definida para exponentes racionales.

Puesto que la derivada es positiva en todo punto, vemos que la función exponencial es estrictamente creciente en $\mathbb{R}$. En particular es inyectiva. Separando los dos primeros términos de la serie vemos que si $x \geq 0$ entonces $1 + x \leq e^x$, luego $\lim\limits_{x\to+\infty} e^x = +\infty$. A su vez esto implica que

$$\lim_{x\to-\infty} e^x = \lim_{x\to+\infty} e^{-x} = \lim_{x\to+\infty} \frac{1}{e^x} = 0.$$

Por el teorema de los valores intermedios, la función exponencial biyecta $\mathbb{R}$ con el intervalo $]0, +\infty[$.

Ejemplo Aplicando n veces la regla de L'Hôpital se concluye claramente que

$$\lim_{x\to+\infty} \frac{x^n}{e^x} = 0,$$

y cambiando x por $1/x$ llegamos a que

$$\lim_{x\to 0^+} \frac{e^{-1/x}}{x^n} = 0, \quad n = 0, 1, 2, \ldots \tag{3.5}$$

Esto implica que la función $h : \mathbb{R} \longrightarrow \mathbb{R}$ dada por

$$h(x) = \begin{cases} e^{-1/x} & \text{si } x > 0 \\ 0 & \text{si } x \leq 0 \end{cases}$$

es de clase C^∞ en $\mathbb{R}$. En efecto, una simple inducción prueba que las derivadas de h para $x > 0$ son de la forma

$$\frac{e^{-1/x}}{x^n} P(x),$$

donde $P(x)$ es un polinomio. De aquí que las derivadas sucesivas de h en 0 existen y valen todas 0. En efecto, admitiendo que existe $h^{k)}(0) = 0$ (para $k \geq 0$) la derivada $h^{k+1)}(0)$ se obtiene por un límite cuando $\Delta x \to 0$ que por la izquierda es claramente 0 y por la derecha es de la forma

$$\lim_{\Delta x \to 0^+} \frac{e^{-1/\Delta x}}{\Delta x^n} P(\Delta x),$$

de modo que el primer factor tiende a 0 por (3.5) y el segundo está acotado en un entorno de 0. Por lo tanto existe $h^{k+1)}(0) = 0$.

En particular, la función $h(x^2)$ es de clase C^∞, sus derivadas son todas nulas en 0 pero es no nula en todo punto distinto de 0. Tenemos así un ejemplo de función de clase C^∞ cuya serie de Taylor en 0 sólo converge (a ella) en 0. ■

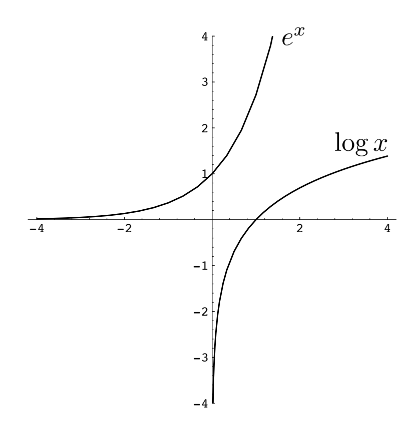
La función h del ejemplo anterior permite construir una familia de funciones que nos serán de gran utilidad más adelante:

Teorema 3.32 *Dados números reales $0 \leq a < b$ existe una función $f : \mathbb{R} \longrightarrow \mathbb{R}$ de clase C^∞ tal que $f(x) > 0$ si $x \in \,]a, b[$ y $f(x) = 0$ en caso contrario.*

DEMOSTRACIÓN: En efecto, la función $h_1(x) = h(x - a)$ se anula sólo en los puntos $x \leq a$ y la función $h_2(b - x)$ se anula sólo en los puntos $x \geq b$. Su producto se anula sólo en los puntos exteriores al intervalo $x \in \,]a, b[$. ■

Definición 3.33 Llamaremos *función logarítmica* a la inversa de la función exponencial, $\log : \,]0, +\infty[\longrightarrow \mathbb{R}$.

He aquí las gráficas de las funciones exponencial y logarítmica.



El teorema de la función inversa nos da que $y = \log x$ es derivable, y su derivada es $y' = 1/(e^y)' = 1/e^y = 1/x$.

De las propiedades de la función exponencial se deducen inmediatamente las de la función logarítmica. Obviamente es una función estrictamente creciente, además verifica la ecuación funcional $\log(xy) = \log x + \log y$. También es claro que $\log 1 = 0$, $\log e = 1$ y

$$\lim_{x\to+\infty} \log x = +\infty, \quad \lim_{x\to 0} \log x = -\infty.$$

Veamos cómo puede calcularse en la práctica un logaritmo. Es decir, vamos a calcular el desarrollo de Taylor de la función log. Obviamente no hay un desarrollo en serie sobre todo $]0,+\infty[$. Si desarrollamos alrededor del 1 a lo sumo podemos obtener una serie convergente en $]0,2[$.

Como las series de potencias centradas en 0 son más fáciles de manejar, vamos a desarrollar en 0 la función $\log(1+x)$. Sus derivadas son

$$(1+x)^{-1}, \quad -(1+x)^{-2}, \quad 2(1+x)^{-3}, \quad -2\cdot 3(1+x)^{-4}, \ldots$$

y en general, la derivada n-sima es $(-1)^{n+1}(n-1)!(1+x)^{-n}$. Puesto que $\log(1+0) = 0$, la serie de Taylor queda:

$$\sum_{n=1}^{\infty} \frac{(-1)^{n+1}}{n} x^n.$$

Como

$$\lim_n \frac{1/(n+1)}{1/n} = \lim_n \frac{n}{n+1} = 1,$$

el radio de convergencia es 1 (como era de esperar), luego la serie converge en $]-1,1[$. De hecho en $x = -1$ obtenemos una serie divergente, pero en $x = 1$ queda $\sum\limits_{n=1}^{\infty} \frac{(-1)^{n+1}}{n}$, que es convergente luego, con exactitud, la serie converge en el intervalo $]-1,1]$.

Según hemos visto en la sección anterior, la función definida por la serie es derivable, y su derivada se obtiene derivando cada monomio, es decir, se trata de la serie geométrica

$$\sum_{n=1}^{\infty}(-1)^{n+1}x^{n-1} = \sum_{n=0}^{\infty}(-x)^n = \frac{1}{1+x}.$$

Resulta, pues, que la serie de Taylor y la función $\log(1+x)$ tienen la misma derivada en $]-1,1[$. Por lo tanto la diferencia entre ambas funciones es una constante, pero como ambas toman el valor 0 en 0, se concluye que

$$\log(1+x) = \sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n}x^n$$

para todo número $x \in\,]-1,1[$.

Ejercicio: Estudiando el resto de Taylor de la función $\log(1+x)$, probar que

$$\sum_{n=1}^{\infty}\frac{(-1)^{n+1}}{n} = \log 2.$$

Para calcular el logaritmo de un número $x > 2$ podemos usar la relación $\log(1/x) = -\log x$.

Ahora podemos definir a^x para cualquier base $a > 0$. La forma más fácil de hacerlo es la siguiente:

Definición 3.34 Sea $a > 0$ y $x \in \mathbb{R}$. Definimos $a^x = e^{x\log a}$.

Notar que, como $\log e = 1$, en el caso $a = e$ la exponencial que acabamos de definir coincide con la que ya teníamos definida. Sin embargo la función e^x se diferencia de las otras exponenciales en que está definida sobre todo el plano complejo y no sólo sobre la recta real. Más adelante interpretaremos esta extensión compleja. Las funciones exponenciales verifican las propiedades siguientes:

$$\begin{aligned}
a^{x+y} &= e^{(x+y)\log a} = e^{x\log a}e^{y\log a} = a^x a^y,\\
\log a^x &= \log e^{x\log a} = x\log a,\\
(a^x)^y &= e^{y\log a^x} = e^{xy\log a} = a^{xy}\\
a^0 &= 1, \quad a^1 = a, \quad a^{-x} = 1/a^x.
\end{aligned}$$

Así mismo es claro que a^x coincide con la exponenciación usual cuando x es un número entero y que sobre números racionales es $a^{p/q} = \sqrt[q]{a^p}$. La función y^x considerada como función de dos variables en $]0,+\infty[\times\mathbb{R}$ es continua.

La derivada de a^x es $(\log a)a^x$, la derivada de x^b es

$$e^{b\log x}\frac{b}{x} = \frac{1}{x}bx^b = bx^{b-1}.$$

Finalmente, puesto que la derivada de a^x es siempre positiva si $a > 1$ y siempre negativa si $a < 1$, tenemos que a^x es monótona y biyecta $\mathbb{R}$ con el intervalo $]0, +\infty[$. Por lo tanto tiene una inversa, que representaremos por $\log_a x$ y se llama *logaritmo* en base a de x. Las propiedades algebraicas de estos logaritmos son las mismas que las de la función log y se demuestran igual. A estas hay que añadir las siguientes, ambas elementales:

$$\log_a x^b = b \log_a x, \quad \log_b x = \frac{\log_a x}{\log_a b}.$$

En particular

$$\log_a x = \frac{\log x}{\log a}.$$

Hay una caracterización importante del número e:

Teorema 3.35 *Se cumple*

$$e = \lim_{x \to +\infty} \left(1 + \frac{1}{x}\right)^x.$$

Demostración: Por definición

$$\left(1 + \frac{1}{x}\right)^x = e^{x \log(1+1/x)},$$

luego basta probar que

$$\lim_{x \to +\infty} x \log\left(1 + \frac{1}{x}\right) = 1.$$

Pasamos la x al denominador como $1/x$ y aplicamos la regla de L'Hôpital, con lo que el límite se transforma en

$$\lim_{x \to +\infty} \frac{\frac{-x^{-2}}{1+1/x}}{-x^{-2}} = \lim_{x \to +\infty} \frac{1}{1 + 1/x} = 1.$$

■

Ejercicio: Probar que, para todo $x \in \mathbb{R}$,

$$e^x = \lim_n \left(1 + \frac{x}{n}\right)^n.$$

Terminamos la sección con una aplicación de los logaritmos junto a técnicas analíticas.

Ejemplo Se define la *media aritmética* de n números reales $x_1, \ldots, x_n$ como $(x_1 + \cdots + x_n)/n$. Si son mayores o iguales que 0 se define su *media geométrica* como $\sqrt[n]{x_1 \cdots x_n}$. Vamos a probar que la media geométrica siempre es menor o igual que la media geométrica. A su vez, deduciremos esto de la desigualdad $\log t \leq t - 1$, válida para todo $t > 0$.

Para probar esta desigualdad vemos que la derivada de $f(t) = t - 1 - \log t$ es $1 - 1/t$, que es negativa si $t < 1$ y positiva si $t > 1$. Por el teorema 3.14, f es decreciente en $]0, 1]$ y creciente en $[1, +\infty[$. Puesto que $f(1) = 0$, es claro entonces que $f(t) \geq 0$ para todo $t > 0$.

Respecto a la desigualdad entre las medias, si uno de los números es nulo la media geométrica es nula y el resultado es obvio. Supongamos que son todos no nulos y sea $x = x_1 \cdots x_n$. Entonces

$$\frac{x_i}{\sqrt[n]{x}} - 1 \geq \log \frac{x_i}{\sqrt[n]{x}},$$

luego sumando obtenemos

$$\frac{\sum\limits_{i=1}^{n} x_i}{\sqrt[n]{x}} - n \geq \log \frac{x_1 \cdots x_n}{x} = 0,$$

con lo que

$$\frac{\sum\limits_{i=1}^{n} x_i}{n} \geq \sqrt[n]{x}.$$

■

3.8 Las funciones trigonométricas

En geometría se definen varias funciones de interés, entre las que destacan las funciones seno y coseno. Si llamamos R a la medida de un ángulo recto, entonces la función $\operatorname{sen} x$ está definida sobre $\mathbb{R}$ y tiene periodo $4R$, es decir, $\operatorname{sen}(x + 4R) = \operatorname{sen} x$ para todo $x \in \mathbb{R}$. Así definido, el valor de R es arbitrario, pues podemos tomar cualquier número real como medida de un ángulo recto. Si queremos que existan ángulos unitarios deberemos exigir que $R > 1/4$. Por ejemplo, si tomamos como unidad de ángulo el grado sexagesimal, entonces $R = 90$. Supongamos que las funciones seno y coseno son derivables así como sus propiedades algebraicas y vamos a calcular su serie de Taylor. Con ello obtendremos una definición analítica alternativa.

En primer lugar, como el coseno tiene un máximo en 0, ha de ser $\cos' 0 = 0$. En cualquier otro punto tenemos

$$\begin{aligned}
\cos' x &= \lim_{h \to 0} \frac{\cos(x+h) - \cos x}{h} = \lim_{h \to 0} \frac{\cos x \cos h - \operatorname{sen} x \operatorname{sen} h - \cos x}{h} \\
&= \cos x \lim_{h \to 0} \frac{\cos h - 1}{h} - \operatorname{sen} x \lim_{h \to 0} \frac{\operatorname{sen} h}{h} \\
&= \cos x \cos' 0 - \operatorname{sen} x \operatorname{sen}' 0 = -\operatorname{sen} x \operatorname{sen}' 0.
\end{aligned}$$

Si llamamos $k = \operatorname{sen}' 0$ concluimos que $\cos' x = -k \operatorname{sen} x$, y similarmente llegamos a que $\operatorname{sen}' x = k \cos x$.

Del mismo modo que hicimos con la exponencial, podemos normalizar las funciones seno y coseno cambiándolas por $\operatorname{sen}(x/k)$ y $\cos(x/k)$. Geométricamente esto significa fijar una unidad de ángulos. Entonces tenemos $\operatorname{sen}' x = \cos x$ y $\cos' x = -\operatorname{sen} x$.

Vemos entonces que las funciones seno y coseno son infinitamente derivables, y sus derivadas están uniformemente acotadas por 1, luego las series de Taylor deben converger en $\mathbb{R}$ a las funciones respectivas. Puesto que $\operatorname{sen} 0 = 0$ y $\cos 0 = 1$, las series han de ser las que consideramos en la definición siguiente:

Definición 3.36 Llamaremos *seno* y *coseno* a las funciones definidas por las series de potencias

$$\operatorname{sen} z = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)!} z^{2n+1}, \quad \cos z = \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n)!} z^{2n}.$$

No es difícil probar directamente la convergencia de estas series sobre todo el plano complejo. El teorema siguiente muestra una sorprendente conexión entre las funciones trigonométricas así definidas y la función exponencial. Observar que la prueba contiene otra demostración alternativa de la convergencia de estas series.

Teorema 3.37 *Para todo $z \in \mathbb{C}$ se cumple*

$$\operatorname{sen} z = \frac{e^{iz} - e^{-iz}}{2i}, \quad \cos z = \frac{e^{iz} + e^{-iz}}{2}.$$

DEMOSTRACIÓN:

$$\frac{e^{iz} + e^{-iz}}{2} = \frac{\sum_{n=0}^{\infty} i^n \frac{z^n}{n!} + \sum_{n=0}^{\infty} (-i)^n \frac{z^n}{n!}}{2} = \sum_{n=0}^{\infty} \frac{i^n + (-i)^n}{2} \frac{z^n}{n!}.$$

Ahora bien, la sucesión $(i^n + (-i)^n)/2$ es simplemente $1, 0, -1, 0, 1, 0, -1, 0 \ldots$, luego queda la serie del coseno. Similarmente se razona con el seno. ■

Derivando término a término las series de Taylor se concluye fácilmente que

$$\operatorname{sen}' x = \cos x, \quad \cos' x = -\operatorname{sen} x.$$

Las fórmulas siguientes son todas consecuencias sencillas del teorema anterior:

$$\operatorname{sen}^2 z + \cos^2 z = 1$$

$$\begin{aligned} \operatorname{sen}(x+y) &= \operatorname{sen} x \cos y + \cos x \operatorname{sen} y, \\ \cos(x+y) &= \cos x \cos y - \operatorname{sen} x \operatorname{sen} y, \\ e^{iz} &= \cos z + i \operatorname{sen} z, \end{aligned}$$

La primera fórmula implica que si $x \in \mathbb{R}$ entonces $-1 \leq \operatorname{sen} x, \cos x \leq 1$. De la última se sigue que para todo $x, y \in \mathbb{R}$ se cumple

$$e^{x+iy} = e^x(\cos y + i \operatorname{sen} y),$$

con lo que tenemos descrita la exponencial compleja en términos de la exponencial real y de las funciones seno y coseno reales.

El hecho de que $\operatorname{sen}' 0 = \cos 0 = 1$ equivale a

$$\lim_{x \to 0} \frac{\operatorname{sen} x}{x} = 1,$$

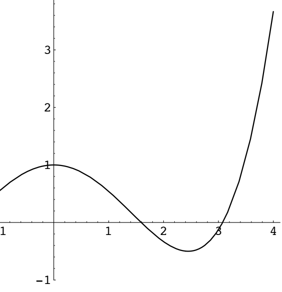
que es otra propiedad del seno que conviene recordar.

Vamos a probar ahora la periodicidad de las funciones trigonométricas reales. El punto más delicado es demostrar que $\cos x$ se anula en algún $x \neq 0$. Para ello probaremos que el coseno es menor o igual que los cuatro primeros términos de su serie de Taylor:

$$\cos x \leq 1 - \frac{x^2}{2} + \frac{x^4}{24}.$$

Esto equivale a probar que $1 - x^2/2 + x^4/24 - \cos x \geq 0$ para $x \geq 0$. Puesto que esta función vale 0 en 0, basta probar que su derivada es positiva. Dicha derivada es $\operatorname{sen} x - x + x^3/6$. Esta función vale también 0 en 0, luego para probar que es positiva (para $x \geq 0$) basta ver que su derivada lo es. Dicha derivada es $\cos x - 1 + x^2/2$. Por el mismo argumento derivamos una vez más y obtenemos $x - \operatorname{sen} x$. Al derivar una vez más llegamos a $1 - \cos x$, que sabemos que es positiva.

Vemos que la gráfica del polinomio de Taylor de grado 4 en 0 de $\cos x$ toma valores negativos. De hecho un simple cálculo nos da que en $\sqrt{3}$ toma el valor $-1/8$, luego $\cos \sqrt{3} \leq -1/8$. Como $\cos 0 = 1$, por continuidad existe un punto $0 < x < \sqrt{3}$ tal que $\cos x = 0$.



Sea $A = \{x > 0 \mid \cos x = 0\}$. El conjunto A es la antiimagen de $\{0\}$ por la aplicación coseno restringida a $[0, +\infty[$. Como $\{0\}$ es cerrado y cos es continua, A es un cerrado. El ínfimo de un conjunto está en su clausura, luego $F = \operatorname{ínf} A \in A$ y así $\cos F = 0$. Es obvio que $F \geq 0$, y como $\cos 0 \neq 0$, ha de ser $0 < F < \sqrt{3}$.

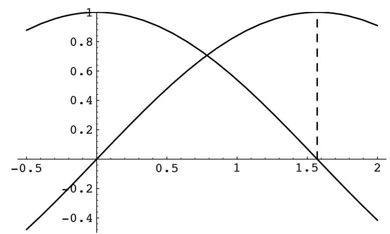
Es costumbre llamar $\pi = 2F$. Así, se cumple $0 < \pi < 2\sqrt{3}$, $\cos(\pi/2) = 0$, pero $\cos x > 0$ en el intervalo $[0, \pi/2[$.

Ahora, como el coseno es la derivada del seno, resulta que $\operatorname{sen} x$ es estrictamente creciente en el intervalo $[0, \pi/2]$. Como $\operatorname{sen} 0 = 0$, resulta que $\operatorname{sen} x \geq 0$ en $[0, \pi/2]$. Concretamente, $\operatorname{sen}(\pi/2)$ es un número positivo que cumple

$$\operatorname{sen}^2(\pi/2) + \cos^2(\pi/2) = \operatorname{sen}^2(\pi/2) + 0 = 1,$$

luego $\operatorname{sen}(\pi/2) = 1$.

Además, $\cos' x = -\operatorname{sen} x \leq 0$ en $[0, \pi/2]$, luego el coseno es estrictamente decreciente en $[0, \pi/2]$. En resumen, tenemos demostrado lo que refleja la gráfica siguiente:

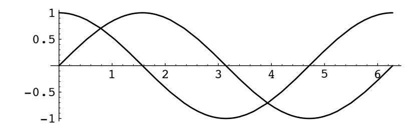


Incidentalmente hemos probado una desigualdad que a veces es de interés: si $x \geq 0$ entonces $\operatorname{sen} x \leq x$. Más en general, $|\operatorname{sen} x| \leq |x|$.

El comportamiento de las funciones seno y coseno fuera del intervalo $[0, \pi/2]$ se deduce de las relaciones trigonométricas que ya hemos probado. Por ejemplo, expresando $\pi = \pi/2 + \pi/2$ obtenemos $\operatorname{sen} \pi = 0$, $\cos \pi = -1$, y a su vez de aquí $\operatorname{sen} 2\pi = 0$, $\cos 2\pi = 1$. Ahora

$$\operatorname{sen}(x + 2\pi) = \operatorname{sen} x, \quad \cos(x + 2\pi) = \cos x,$$

lo que prueba que ambas funciones son periódicas y basta estudiarlas en el intervalo $[0, 2\pi]$. Dejamos a cargo del lector completar la descripción de estas funciones. A partir de lo que ya hemos probado es fácil obtener todos los resultados que se demuestran en geometría. Nos limitaremos a mostrar sus gráficas en $[0, 2\pi]$:



El teorema siguiente se demuestra de forma más natural en geometría, pero vamos a probarlo para tener una construcción completamente analítica de las funciones trigonométricas:

Teorema 3.38 *Sea $z \in \mathbb{C}$, $z \neq 0$ y sea $a \in \mathbb{R}$. Entonces existe un único número real $\theta \in [a, a + 2\pi[$ tal que $z = |z|e^{i\theta} = |z|(\cos\theta + i \operatorname{sen}\theta)$.*

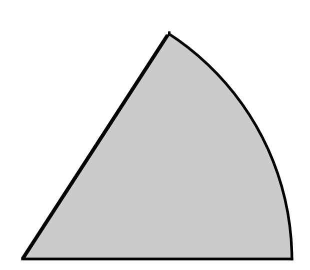
DEMOSTRACIÓN: Sea $z/|z| = x + iy$. Entonces $x^2 + y^2 = 1$. Distinguimos cuatro casos: según el signo de x e y. Todos son análogos, así que supondremos por ejemplo $x \leq 0$, $y \geq 0$. Más concretamente tenemos $-1 \leq x \leq 0$. En el intervalo $[\pi, 3\pi/2]$ se cumple $\cos \pi = -1$, $\cos 3\pi/2 = 0$, luego por continuidad existe un número $\phi \in [\pi, 3\pi/2]$ tal que $\cos \phi = x$. Entonces $1 = x^2 + y^2 = \cos^2 \phi + \operatorname{sen}^2 \phi$, por lo que $y^2 = \operatorname{sen}^2 \phi$ y, como ambos son negativos, ha de ser $y = \operatorname{sen} \phi$. Así pues, $z = |z|(\cos \phi + i \operatorname{sen} \phi) = |z|e^{i\phi}$.

Existe un número entero p tal que $\theta = 2p\pi + \phi \in [a, a + 2\pi[$. Entonces, teniendo en cuenta que $e^{2p\pi i} = 1$, resulta que $z = |z|e^{i\phi}e^{2p\pi i} = |z|e^{i\theta}$.

La unicidad se debe a que si $|z|e^{i\theta_1} = |z|e^{i\theta_2}$, entonces $e^{i(\theta_1-\theta_2)} = 1$, luego $\cos(\theta_1 - \theta_2) = 1$ y $\operatorname{sen}(\theta_1 - \theta_2) = 0$, ahora bien, $\cos x = 1$ y $senx = 0$ sólo ocurre en $x = 0$ en el intervalo $[0, 2\pi[$, luego sólo ocurre en los números reales de la forma $2k\pi$, con $k \in \mathbb{Z}$. Así pues $\theta_1 - \theta_2 = 2k\pi$, y si ambos están en el intervalo $[a, a + 2\pi[$, ha de ser $\theta_1 = \theta_2$. ■

Un *argumento* de un número complejo $z \neq 0$ es un número real θ tal que $z = |z|e^{i\theta}$. Hemos probado que cada número complejo no nulo tiene un único argumento en cada intervalo $[a, a + 2\pi[$. En particular en el intervalo $[0, 2\pi[$.

Recordemos que para conseguir que la derivada del seno fuera el coseno hemos tenido que fijar una medida de ángulos concreta. El ángulo de medida 1 respecto a esta unidad, es decir, el ángulo que forman los vectores $(1, 0)$ y $(\cos 1, \operatorname{sen} 1)$, recibe el nombre de *radián*[1] La figura muestra un radián.



Las funciones seno y coseno nos permiten mostrar algunos ejemplos de interés sobre derivabilidad.

Ejemplo La función $f : \mathbb{R} \longrightarrow \mathbb{R}$ dada por

$$f(x) = \begin{cases} x^2 \operatorname{sen} \frac{1}{x} & \text{si } x \neq 0 \\ 0 & \text{si } x = 0 \end{cases}$$

es derivable en $\mathbb{R}$. El único punto donde esto no es evidente es $x = 0$, pero

$$f'(0) = \lim_{h \to 0} h \operatorname{sen} \frac{1}{h} = 0.$$

Para probar esto observamos en general que el producto de una función acotada por otra que tiende a 0 tiende a 0 (basta aplicar la definición de límite).

Sin embargo la derivada no es continua, pues en puntos distintos de 0 vale

$$f'(x) = 2x \operatorname{sen} \frac{1}{x} - \cos \frac{1}{x},$$

y es fácil ver que el primer sumando tiende a 0 en 0 (como antes), mientras que el segundo no tiene límite, luego no existe $\lim_{x \to 0} f'(x)$.

Los mismos cálculos que acabamos de realizar prueban una limitación de la regla de L'Hôpital. Consideremos la función $g(x) = x$. Entonces

$$\lim_{x \to 0} \frac{f(x)}{g(x)} = 0,$$

[1]El nombre se debe a que, según probaremos en el capítulo siguiente, si un arco de circunferencia mide un radián, entonces su longitud es igual al radio. Sería más adecuado llamarlo ángulo “radiante”.

pero si intentamos calcular el límite por la regla de L'Hôpital nos encontramos con

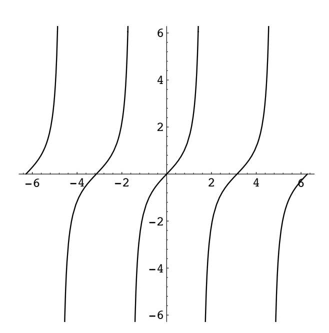
$$\lim_{x\to 0} \frac{f'(x)}{g'(x)} = \lim_{x\to 0} f'(x),$$

y ya hemos visto que este límite no existe. La regla de L'Hôpital sólo afirma que si existe el límite del cociente de derivadas también existe el límite original y ambos coinciden, pero es importante recordar que si el segundo límite no existe de ahí no podemos deducir que el primero tampoco exista. ∎

Otra función trigonométrica importante es la tangente, definida como

$$\tan z = \frac{\operatorname{sen} z}{\cos z}.$$

No es difícil probar que la función coseno se anula únicamente sobre los múltiplos enteros de $\pi/2$ (no tiene ceros imaginarios). En efecto, si $\cos z = 0$, por definición $e^{iz}+e^{-iz} = 0$, luego $e^{2iz} = -1$. Si $z = a+bi$, queda $e^{2ia}e^{-2b} = -1$ y tomando módulos, $e^{-2b} = 1$, luego $b = 0$ y z es real. Igualmente ocurre con el seno. Por lo tanto la tangente está definida sobre todos los números complejos que no son múltiplos enteros de $\pi/2$. Claramente su restricción a $\mathbb{R}$ es derivable en su dominio, y su derivada es $1/\cos^2 x = 1 + \tan^2 x$. En particular es siempre positiva, luego la tangente es creciente. Su gráfica es:



Es fácil ver que la función tangente biyecta el intervalo $]-\pi/2, \pi/2[$ con la recta real. Junto con ésta, tenemos también las biyecciones siguientes:

$$\operatorname{sen} : \left[-\frac{\pi}{2}, \frac{\pi}{2}\right] \longrightarrow [-1, 1], \quad \cos : [0, \pi] \longrightarrow [-1, 1].$$

Por lo tanto podemos definir las funciones inversas, llamadas respectivamente, *arco seno, arco coseno* y *arco tangente:*

$$\operatorname{arcsen} : [-1, 1] \longrightarrow \left[-\frac{\pi}{2}, \frac{\pi}{2}\right], \operatorname{arccos} : [-1, 1] \longrightarrow [0, \pi],$$

$$\arctan : \mathbb{R} \longrightarrow \left]-\pi/2, \pi/2\right[.$$

El teorema de la función inversa permite calcular sus derivadas:

$$\operatorname{arcsen}' x = \frac{1}{\sqrt{1-x^2}}, \quad \operatorname{arccos}' x = \frac{-1}{\sqrt{1-x^2}}, \quad \arctan' x = \frac{1}{1+x^2}.$$

Por ejemplo, si $y = \operatorname{arcsen} x$, entonces $x = \operatorname{sen} y$, luego $\frac{dx}{dy} = \cos y$, luego

$$\frac{dy}{dx} = \frac{1}{\cos y} = \frac{1}{\sqrt{1 - \operatorname{sen}^2 y}} = \frac{1}{\sqrt{1 - x^2}}.$$

Vamos a calcular la serie de Taylor de la función arco tangente. No podemos calcular directamente las derivadas, pues las expresiones que se obtienen son cada vez más complicadas y no permiten obtener una fórmula general. En su lugar emplearemos la misma técnica que hemos usado en la sección anterior para calcular la serie del logaritmo. Claramente

$$\frac{1}{1 + x^2} = \frac{1}{1 - (-x^2)} = \sum_{n=0}^{\infty} (-1)^n x^{2n}, \quad \text{para } |x| < 1.$$

Ahora es fácil obtener una serie cuya derivada sea la serie anterior, a saber:

$$\sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1}.$$

Es fácil ver que su radio de convergencia es 1.

Por lo tanto esta serie se diferencia en una constante de la función $\arctan x$ en el intervalo $]-1, 1[$, pero como ambas funciones toman el valor 0 en $x = 0$, concluimos que son iguales, o sea:

$$\arctan x = \sum_{n=0}^{\infty} \frac{(-1)^n}{2n+1} x^{2n+1}, \quad \text{para } |x| < 1.$$

Cuando $x = \pm 1$ la serie se convierte en una serie alternada cuyo término general es decreciente y tiende a 0, luego por el criterio de Leibniz también converge. No vamos a demostrarlo aquí (ver la pág. 237), pero el límite resulta ser $\arctan(\pm 1)$. Puesto que $\arctan 1 = \pi/4$, esto nos lleva a la conocida fórmula de Leibniz para el cálculo de π:

$$\pi = 4\left(1 - \frac{1}{3} + \frac{1}{5} - \frac{1}{7} + \frac{1}{9} - \frac{1}{11} + \cdots\right).$$

Esta fórmula converge muy lentamente a π, en el sentido de que es necesario calcular muchos términos para obtener pocas cifras exactas. Hay otras expresiones más complicadas pero más eficientes. Veamos una de ellas. Es fácil probar la fórmula de la tangente del ángulo doble:

$$\tan 2\alpha = \frac{2 \tan \alpha}{1 - \tan^2 \alpha}.$$

De aquí se sigue que

$$\tan\left(2 \arctan \frac{1}{5}\right) = \frac{5}{12}, \quad \tan\left(4 \arctan \frac{1}{5}\right) = \frac{120}{119}.$$

Teniendo en cuenta que $\arctan(-x) = -\arctan x$ resulta

$$\tan\left(4\arctan\frac{1}{5} - \arctan\frac{1}{239}\right) = \frac{\frac{120}{119} - \frac{1}{239}}{1 + \frac{120}{119}\frac{1}{239}} = 1,$$

con lo que, finalmente,

$$\pi = 4\left(4\arctan\frac{1}{5} - \arctan\frac{1}{239}\right),$$

o sea,

$$\pi = \sum_{n=0}^{\infty}\frac{(-1)^n 4}{2n+1}\left(\frac{4}{5^{2n+1}} - \frac{1}{239^{2n+1}}\right).$$

Esta serie converge muy rápidamente. Veamos sus primeras sumas parciales:

0	**3,1**8326359832635983263598326360
1	**3,14**059702932606031430453110658
2	**3,141**62102932503442504683251712
3	**3,14159**177218217729501821229111
4	**3,1415926**8240439951724025983607
5	**3,14159265**261530860814935074767
6	**3,141592653**62355476199550459382
7	**3,1415926535**8860222866217126049
8	**3,141592653589**83584748570067225
9	**3,14159265358979**169691727961962
10	**3,1415926535897932**9474737485772
11	**3,14159265358979323**639184094467
12	**3,141592653589793238**53932459267
13	**3,1415926535897932384**5978816126
14	**3,141592653589793238462**75020768
15	**3,1415926535897932384626**3936981

Vemos que con 6 sumas tenemos ya una aproximación con diez cifras exactas, y con 15 superamos las 20 cifras.

***Las funciones hiperbólicas** El lector que esté familiarizado con la geometría hiperbólica habrá notado la similitud formal entre las fórmulas del teorema 3.37 y las definiciones de las razones trigonométricas hiperbólicas. Esta similitud se traduce en las relaciones siguientes:

$$\begin{aligned} \operatorname{senh} x &= \frac{e^x - e^{-x}}{2} = \frac{e^{i(-ix)} - e^{-i(-ix)}}{2} = i\operatorname{sen}(-ix) = \frac{\operatorname{sen} ix}{i}, \\ \cosh x &= \frac{e^x + e^{-x}}{2} = \frac{e^{i(-ix)} + e^{-i(-ix)}}{2} = \cos(-ix) = \cos ix. \end{aligned}$$

En definitiva, tenemos

$$\begin{aligned} \operatorname{sen}(ix) &= i\operatorname{senh} x, \\ \cos(ix) &= \cosh x, \\ \tan(ix) &= i\tanh x. \end{aligned}$$

De aquí se siguen, por ejemplo, las relaciones $\cosh^2 x - \operatorname{senh}^2 x = 1$, o las fórmulas del seno y el coseno de una suma, etc.

Ejercicio: Deducir de las relaciones anteriores las series de Taylor de las funciones $\operatorname{senh} x$ y $\cosh x$.

A partir de las definiciones es fácil probar

$$\operatorname{senh}' x = \cosh x, \quad \cosh' x = \operatorname{senh} x, \quad \tanh' x = \frac{1}{\cosh^2 x} 1 - \tanh^2 x.$$

Puesto que $\cosh x \geq 0$, el seno hiperbólico es creciente. De hecho es una función biyectiva, como puede probarse a partir de la propia definición: Si $x = \operatorname{senh} y$ entonces

$$e^y - e^{-y} - 2x = 0, \quad \Rightarrow \quad e^{2y} - 2xe^y - 1 = 0, \quad \Rightarrow e^y = x + \sqrt{x^2 + 1},$$

luego la inversa del seno hiperbólico es la función *argumento del seno hiperbólico* dada por

$$\operatorname{arg\,senh} x = \log\bigl(x + \sqrt{x^2 + 1}\,\bigr).$$

Teniendo en cuenta su derivada, el coseno hiperbólico es decreciente en $]-\infty, 0]$ y creciente en $[0, +\infty[$. Como $\cosh 0 = 1$, la imagen está en $[1, +\infty[$. Existe la inversa $\operatorname{arg\,cosh} : [1, +\infty[\longrightarrow [0, +\infty[$ dada por

$$\operatorname{arg\,cosh} x = \log\bigl(x + \sqrt{x^2 - 1}\,\bigr).$$

Teniendo en cuenta las relaciones

$$\tan^2 x = 1 - \frac{1}{\cosh^2 x}, \quad \tanh(-x) = -\tanh x,$$

es claro que $\tanh : \mathbb{R} \longrightarrow]-1, 1[$ es biyectiva, luego tiene como inversa a la función $\operatorname{arg\,tanh} :]-1, 1[\longrightarrow \mathbb{R}$.

Dejamos a cargo del lector el cálculo de las derivadas de los argumentos hiperbólicos. ■

***Geometrías no euclídeas** Los resultados de esta sección nos permiten mostrar una conexión importante entre la geometría euclídea y las geometrías elíptica e hiperbólica. Antes de entrar en ello, observemos que en la geometría euclídea existe una asimetría entre la medida de longitudes y la medida de ángulos. En efecto, no es posible definir geométricamente una unidad de longitud. El metro se define actualmente en términos del segundo y de la velocidad de la luz, y a su vez el segundo se define en términos de una propiedad física del átomo de kriptón; antiguamente el metro se definía como la longitud del patrón de platino que se hallaba en París. En cualquier caso, si alguien quiere construir un metro con precisión se ve obligado a observar átomos de kriptón, a viajar a París o algo similar. En cambio, no es necesario viajar a París para construir un ángulo recto, o un radián, o un grado sexagesimal con precisión.

Podría haber un ángulo patrón en París, pero no es necesario porque existen unidades naturales de ángulo, en el sentido de que pueden definirse por medios puramente geométricos. Nosotros hemos definido el radián y no hemos tenido que aludir a ningún átomo.

Esta asimetría no se da en las geometrías no euclídeas. Pensemos por ejemplo en la geometría elíptica. Los segmentos pueden medirse también en radianes, o en rectos, o en grados sexagesimales. Una longitud de un recto es simplemente la mitad de la longitud de una recta cualquiera. Una longitud de un radián es simplemente $2/\pi$ rectos. Desde un punto de vista más conceptual, esto se refleja en que la geometría euclídea tiene semejanzas que no son isometrías (aplicaciones que conservan ángulos pero no longitudes, de modo que dos segmentos cualesquiera son semejantes), mientras que en las geometrías no euclídeas todas las semejanzas son isometrías.

La ausencia de unidades naturales de longitud (o la presencia de semejanzas) hace que la geometría euclídea sea invariante a escala. La geometría de Liliput es exactamente la misma que la nuestra, y así sería aunque los liliputienses midieran una millonésima de milímetro. No habría ningún argumento objetivo para concluir que ellos son "pequeños" y nosotros "grandes" o "normales". Las nociones de "grande" y "pequeño" son relativas a la unidad de medida, y ésta es arbitraria. No ocurre lo mismo en las geometrías no euclídeas. Supongamos que la superficie de la Tierra fuera completamente esférica, sin hoyos ni elevaciones. Ahora sí tiene sentido decir que un metro es objetivamente "pequeño", con respecto a una unidad natural de longitud, pues un metro es 20 millones de veces más pequeño que un meridiano (una recta elíptica). Quizá el lector piense que esta noción de pequeñez no deja de ser subjetiva. Ciertamente no es rigurosa en el sentido de que no podemos determinar cuándo una longitud deja de ser pequeña, pero es la forma más natural de expresar un hecho objetivo que vamos a explicar a continuación.

Supongamos que trazamos un triángulo en nuestra Tierra perfecta. Digamos que uno de sus lados mide un metro. Entonces podemos aplicarle el teorema del coseno para un lado, es decir, la fórmula:

$$\cos\alpha = -\cos\beta\cos\gamma + \operatorname{sen}\beta\operatorname{sen}\gamma\,\cos a.$$

Ahora bien, si entendemos que las funciones trigonométricas que aparecen son las que hemos estudiado en este capítulo, los ángulos han de estar en radianes. Si el lado a mide un metro, su medida en radianes es aproximadamente $a \approx 1{,}6\cdot 10^{-7}$, y entonces $\cos a \approx 0{,}9999999999999872$. Por consiguiente nuestro triángulo cumple

$$\cos\alpha \approx -\cos\beta\cos\gamma + \operatorname{sen}\beta\operatorname{sen}\gamma = -\cos(\beta+\gamma) = \cos(\pi-\beta-\gamma),$$

de donde se sigue que $\alpha+\beta+\gamma \approx \pi + 2k\pi$. Como la suma de los ángulos de un triángulo elíptico está entre π y 3π, concluimos que la suma ha de parecerse a π o a 3π. Enseguida descartaremos el segundo caso y concluiremos

$$\alpha+\beta+\gamma \approx 2\pi.$$

Si hacemos cálculos concretos veremos que esta aproximación excede nuestra capacidad de discernimiento, por lo que aparentemente nuestro triángulo cumplirá la ley euclídea de que la suma de sus ángulos es igual a π.

La razón por la que la suma de los ángulos no puede acercarse a 3π nos la da el teorema del coseno para un ángulo:

$$\cos a = \cos b \cos c + \operatorname{sen} b \operatorname{sen} c \, \cos \alpha.$$

Si ahora sustituimos $\cos a \approx 1$, $\cos b \approx 1$, $\cos b \approx 1$ obtenemos la relación $\operatorname{sen} b \operatorname{sen} c \, \cos \alpha \approx 0$, que no expresa más que el hecho de que cuando b y c son pequeños $\operatorname{sen} b \approx 0 \approx \operatorname{sen} c$. Vamos a aproximar las razones trigonométricas por sus polinomios de Taylor de grado 2 en 0.

Consideramos grado 2 porque, para el coseno, el polinomio de Taylor de grado 1 es también $P_1(x) = 1$, luego estamos en el mismo caso de antes. En cambio el de grado 2 nos da la aproximación:

$$\cos x \approx 1 - \frac{x^2}{2}.$$

Respecto al seno, el polinomio de grado 1 coincide con el de grado 2. Ambos nos dan la aproximación $\operatorname{sen} x \approx x$.

Ejercicio: Comprobar que para $a \approx 1{,}6 \cdot 10^{-7}$ el resto del polinomio de Taylor de grado 3 (que es el mismo que el de grado 2) del coseno en 0 es menor que $3 \cdot 10^{-29}$. El resto de grado 2 del seno es menor que $7 \cdot 10^{-22}$.

Por consiguiente

$$\cos b \cos c \approx 1 - \frac{b^2}{2} - \frac{c^2}{2} + \frac{b^2 c^2}{4}.$$

Ésta es una aproximación de cuarto grado. Podemos despreciar el último término y quedarnos con

$$\cos b \cos c \approx 1 - \frac{b^2}{2} - \frac{c^2}{2}.$$

Aunque no hemos definido este concepto, lo cierto es que $1 - x^2/2 - y^2/2$ es el polinomio de Taylor de segundo grado de la función de dos variables $\cos x \cos y$. Con estas aproximaciones el teorema del coseno se convierte en

$$1 - \frac{a^2}{2} \approx 1 - \frac{b^2}{2} - \frac{c^2}{2} + bc \, \cos \alpha,$$

o equivalentemente,

$$a^2 \approx b^2 + c^2 - 2bc \cos \alpha,$$

que es el teorema del coseno euclídeo. Esto significa que los ángulos de nuestro triángulo serán aproximadamente los mismos que los del triángulo euclídeo de lados a, b, c, luego su suma se parecerá a π y no a 3π.

En general, todas las fórmulas de la geometría elíptica aproximan a fórmulas euclídeas cuando las longitudes involucradas son pequeñas. La aproximación $\operatorname{sen} x \approx x$ transforma el teorema de los senos elíptico en el correspondiente euclídeo.

Ejercicio: Comprobar que el teorema de Pitágoras elíptico: $\cos a = \cos b \cos c$ se convierte aproximadamente en el euclídeo para triángulos pequeños.

Resulta así que la geometría elíptica a gran escala es muy diferente de la geometría elíptica a pequeña escala, pues ésta última es aproximadamente euclídea. Esto se expresa diciendo que la geometría elíptica es localmente euclídea. La interpretación geométrica es clara: la geometría elíptica es localmente igual a la geometría de una esfera, y la geometría de una esfera se aproxima localmente a la de cualquiera de sus planos tangentes, que es euclídea.

Todo lo dicho vale igualmente para la geometría hiperbólica. En primer lugar hemos de observar que también en esta geometría hay unidades naturales de longitud, definibles geométricamente. Por ejemplo, si un triángulo equilátero tiene todos sus lados unitarios, el teorema del coseno nos permite calcular sus ángulos. Todos miden

$$\alpha = \arccos \frac{e^2+1}{e^2+2e+1} \approx 0{,}92 \text{ rad.}$$

Por lo tanto podemos definir la unidad de longitud hiperbólica como la longitud del lado del triángulo equilátero cuyos ángulos miden α. La selección de esta unidad de longitud se lleva a cabo en el momento en que definimos la distancia entre dos puntos mediante la fórmula

$$d(P,Q) = \frac{1}{2} \log \mathcal{R}(P,Q,Q_\infty,P_\infty).$$

Si cambiamos la base del logaritmo o si cambiamos la constante $1/2$ estamos cambiando de unidad de longitud (ambos cambios son equivalentes).

Para transformar las fórmulas hiperbólicas en fórmulas euclídeas basta usar las aproximaciones de Taylor:

$$\operatorname{senh} x \approx x, \quad \cosh x \approx 1 + \frac{x^2}{2}.$$

Dejamos los detalles a cargo del lector. ■

3.9 Primitivas

El teorema 3.16 afirma que una función derivable está completamente determinada por su derivada y su valor en un punto. A menudo se plantea el problema práctico de determinar una función a partir de estos datos.

Definición 3.39 Diremos que una función $F : I \longrightarrow \mathbb{R}$ definida en un intervalo abierto I es una *primitiva* de una función $f : I \longrightarrow \mathbb{R}$ si F es derivable en I y $F' = f$.

No estamos en condiciones de decir mucho sobre cuándo una función tiene primitiva. Nos limitaremos a hacer algunas observaciones teóricas que en casos concretos nos permitirán encontrar primitivas de funciones dadas. En estos términos, lo que afirma el teorema 3.16 es que si una función f admite una primitiva F, entonces el conjunto de todas las primitivas de f está formado por las funciones de la forma $F + c$, para cada $c \in \mathbb{R}$. Esto hace que, aunque F no esté unívocamente determinada por f, dados a, $b \in I$, el incremento $F(b) - F(a)$ sí lo está. Conviene introducir la notación

$$\big[F(x)\big]_a^b = F(b) - F(a).$$

En el lenguaje del cálculo infinitesimal, si sabemos que $dy = f(x)dx$, esto significa que la función y experimenta un incremento infinitesimal de $f(x)dx$ cada vez que la variable se incrementa en dx. Supongamos que y está definida en $[a, b]$ y conocemos $y(a)$. Entonces $y(b) \approx y(a) + f(a)(b - a)$. En realidad éste es el valor que toma en b la recta tangente a y en a. Obtendremos una aproximación mejor si dividimos el intervalo $[a, b]$ en partes iguales, digamos $a = x_0 < x_1 < \cdots < x_n = b$, todas de longitud Δx, y vamos calculando:

$$\begin{array}{rcl} y(x_0) & = & y(a), \\ y(x_1) & \approx & f(x_0)\Delta x + y(a), \\ y(x_2) & \approx & f(x_0)\Delta x + f(x_1)\Delta x + y(a), \\ \cdots & & \cdots\cdots\cdots\cdots \end{array}$$

En definitiva,

$$y(b) - y(a) \approx \sum_{i=1}^{n} f(x_{i-1})\Delta x$$

La aproximación será mejor cuantas más partes consideremos, es decir, cuanto menor sea Δx. Así, el incremento de la primitiva puede pensarse como una suma de infinitos sumandos correspondientes a un incremento infinitesimal dx. Por ello la notación clásica para el incremento de una primitiva F de una función $f(x)$ en un intervalo $[a, b]$ es

$$\int_a^b f(x)\,dx = \big[F(x)\big]_a^b \tag{3.6}$$

donde el símbolo proviene de una S de "suma" (en realidad del latín *summa*) y se lee *integral* de la función f respecto a x en $[a, b]$ (porque es el incremento "entero" que se obtiene al sumar sus incrementos infinitesimales). Los números a y b se llaman *límites* de la integral. Para convertir al cálculo integral en una teoría matemática eficaz es necesario tener en cuenta estas ideas y tratarlas con rigor, pero de momento no vamos a entrar en ello y vamos a limitarnos a considerar la integración como la operación inversa a la derivación. Así pues, tomamos (3.6) como definición de la integral de f. Notemos que si dejamos el límite superior como variable obtenemos una primitiva concreta G de f, a saber,

la única que cumple $G(a) = 0$:

$$G(t) = \int_a^x f(x)\,dx = F(x) - F(a).$$

Cuando queremos referirnos a una primitiva arbitraria de una función usamos esta misma notación integral, pero omitimos los límites de integración, así,

$$\int f(x)\,dx = F(x) + c, \quad \text{donde } c \in \mathbb{R}.$$

Veremos que esta notación es consistente con los otros usos que venimos haciendo de los símbolos dx. De la propia definición de primitiva se sigue que

$$\int y'(x)\,dx = y(x) + c.$$

Con la notación $dy = y'\,dx$ esto se escribe así:

$$\int dy = y + c.$$

Cada regla de derivación da lugar a una regla de integración. Por ejemplo, el hecho de que la derivada de una suma es la suma de las derivadas implica que una suma tiene por primitiva a la suma de las primitivas. Más en general, dadas dos funciones u, v,

$$\int(\alpha\,u + \beta\,v)\,dx = \alpha\int u\,dx + \beta\int v\,dx, \quad \text{para } \alpha, \beta \in \mathbb{R}.$$

Uniendo esto a la regla evidente:

$$\int x^n = \frac{x^{n+1}}{n+1} + c, \quad n \neq -1,$$

podemos integrar cualquier polinomio. El caso exceptuado es claro:

$$\int x^{-1}\,dx = \log x + c.$$

La regla de derivación del producto requiere más atención: dadas dos funciones derivables u y v tenemos que $d(uv) = udv + vdu$, luego integrando tenemos

$$\int u\,dv = uv - \int v\,du.$$

Esta fórmula se conoce como "regla de integración por partes". Obviamente de aquí se sigue a su vez la versión con límites:

$$\int_a^b u\,dv = [uv]_a^b - \int_a^b v\,du.$$

Ejemplo Vamos a calcular $\int xe^x\,dx$. Para ello llamamos $u = x$ y $dv = e^x\,dx$. Claramente entonces $du = dx$ y $v = \int dv = \int e^x\,dx = e^x$. La fórmula anterior nos da

$$\int xe^x\,dx = xe^x - \int e^x\,dx = xe^x - e^x + c.$$

Observar que hemos omitido la constante al calcular $v = e^x$. Es claro que para aplicar la regla podemos tomar como v una primitiva fija cualquiera. ∎

Ejemplo Sea $I_{m,n} = \int \operatorname{sen}^m x\,\cos^n x\,dx$. Vamos a encontrar unas expresiones recurrentes que nos permitan calcular estas integrales. Integramos por partes tomando $u = \cos^{n-1} x$, $dv = \operatorname{sen}^m x\,\cos x\,dx$. De este modo:

$$\begin{aligned}
I_{m,n} &= \frac{\operatorname{sen}^{m+1} x\,\cos^{n-1} x}{m+1} + \frac{n-1}{m+1}\int \operatorname{sen}^{m+1} x\,\cos^{n-2} x\,\operatorname{sen} x\,dx \\
&= \frac{\operatorname{sen}^{m+1} x\,\cos^{n-1} x}{m+1} + \frac{n-1}{m+1}\int \operatorname{sen}^{m} x\cos^{n-2} x(1-\cos^2 x)\,dx \\
&= \frac{\operatorname{sen}^{m+1} x\,\cos^{n-1} x}{m+1} + \frac{n-1}{m+1}(I_{m,n-2} - I_{m,n}).
\end{aligned}$$

Despejando llegamos a

$$I_{m,n} = \frac{\operatorname{sen}^{m+1} x\cos^{n-1} x}{m+n} + \frac{n-1}{m+n} I_{m,n-2}.$$

Similarmente se prueba

$$I_{m,n} = -\frac{\operatorname{sen}^{m-1} x\cos^{n+1} x}{m+n} + \frac{m-1}{m+n} I_{m-2,n}.$$

Estas fórmulas reducen el cálculo de cualquier integral $I_{m,n}$ al cálculo de las cuatro integrales

$$\int dx, \quad \int \operatorname{sen} x\,dx, \quad \int \cos x\,dx, \quad \int \operatorname{sen} x\cos x\,dx,$$

y todas ellas son inmediatas (para la última aplicamos la fórmula del seno del ángulo doble). ∎

Por último veamos en qué se traduce la regla de la cadena. Supongamos que tenemos una integral $\int u(x)\,dx$, que $F(x)$ es una primitiva de $u(x)$ (la que queremos calcular) y que $x = x(t)$ es una función derivable con derivada no nula (que por consiguiente tiene inversa derivable $t = t(x)$). Entonces por la regla de la cadena $F(x(t))' = F'(x(t))\,x'(t) = u(x(t))x'(t)$, luego

$$\int u(x(t))\,x'(t)\,dt = F(x(t)) + c.$$

Así pues, para calcular $\int u(x)\,dx$ podemos sustituir x por $x(t)$ y dx por $x'(t)\,dt$ y calcular una primitiva $G(t)$ de la función resultante, ésta será $F(x(t)) + c$, luego

si sustituimos $G(t(x)) = F(x(t(x)))+c = F(x)+c$, obtenemos una primitiva de la función original. A esta técnica se la llama *integración por sustitución.* La versión con límites es:

$$\int_{t(a)}^{t(b)} u(x(t))\,x'(t)\,dt = F(x(t(b))) - F(x(t(a))) = F(b) - F(a) = \int_a^b u(x)\,dx,$$

o sea:

$$\int_a^b u(x)\,dx = \int_{t(a)}^{t(b)} u(x(t))\,x'(t)\,dt.$$

Ejemplo Vamos a calcular $\int \sqrt{1-x^2}\,dx$. El integrando está definido en el intervalo $]-1,1[$. Consideramos la función $x = \operatorname{sen} t$, definida y biyectiva en $]0,\pi[$. Entonces $dx = \cos t\,dt$ y se cumple

$$\int \sqrt{1-x^2}\,dx = \int \sqrt{1-\operatorname{sen}^2 t}\cos t\,dt = \int \cos^2 t\,dt = \int \frac{1+\cos 2t}{2}\,dt$$

$$= \frac{1}{2}\int dt + \frac{1}{4}\int 2\cos 2t\,dt = \frac{t}{2} + \frac{1}{4}\operatorname{sen} 2t + c = \frac{t}{2} + \frac{1}{2}\operatorname{sen} t\cos t + c$$

$$= \frac{\operatorname{arcsen} x}{2} + \frac{x\sqrt{1-x^2}}{2} + c.$$

■

En general, si la primitiva de una función en un intervalo $]a,b[$ se extiende continuamente al intervalo $[a,b]$, dicha extensión es obviamente única, por lo que podemos calcular la integral desde a hasta b, que se interpreta como el incremento completo de la primitiva. Así, en el ejemplo anterior tenemos

$$\int_{-1}^{1} \sqrt{1-x^2}\,dx = \frac{\pi}{2}.$$

3.10 Apéndice: La trascendencia de e y π

Para acabar el capítulo probaremos que las constantes e y π que nos han aparecido son números trascendentes (sobre $\mathbb{Q}$). Supondremos al lector familiarizado con la teoría de números algebraicos. Aunque es muy sencillo probar que casi todos los números reales son trascendentes, pues el conjunto de números algebraicos es numerable y $\mathbb{R}$ no lo es. No es fácil, en cambio, probar la trascendencia de un número particular. Hermite fue el primero en demostrar la trascendencia de e y Lindemann probó después la de π. Aquí veremos unas pruebas más sencillas, pero hay que señalar que la prueba de Lindemann se generaliza a un teorema más potente, en virtud del cual los valores que toman las funciones $\operatorname{sen} x$, $\cos x$, e^x, $\log x$, etc. sobre números algebraicos son—salvo los casos triviales— números trascendentes. Por ello a estas funciones se les llama también *funciones trascendentes.*

Comenzamos introduciendo unos convenios útiles de notación:

Definición 3.40 Escribiremos $h^r = r!$, de modo que si $f(z) = \sum_{r=0}^{m} c_r z^r \in \mathbb{C}[z]$, entonces $f(h)$ representará

$$f(h) = \sum_{r=0}^{m} c_r h^r = \sum_{r=0}^{m} c_r r!$$

Igualmente, $f(z+h)$ será el polinomio que resulta de sustituir y^r por $h^r = r!$ en la expresión desarrollada de $f(z+y)$. Concretamente:

$$f(z+y) = \sum_{r=0}^{m} c_r (z+y)^r = \sum_{r=0}^{m} c_r \sum_{k=0}^{r} \binom{r}{k} z^{r-k} y^k = \sum_{k=0}^{m} \left(\sum_{r=k}^{m} c_r \binom{r}{k} z^{r-k} \right) y^k,$$

luego

$$f(z+h) = \sum_{k=0}^{m} \left(\sum_{r=k}^{m} c_r \binom{r}{k} z^{r-k} \right) k! = \sum_{k=0}^{m} \left(\sum_{r=k}^{m} \frac{r!}{(r-k)!} c_r z^{r-k} \right) = \sum_{k=0}^{m} f^{k)}(z),$$

donde $f^{k)}(z)$ es la k-ésima derivada formal del polinomio f. Observar que

$$f(z+h) = \sum_{r=0}^{m} c_r (z+h)^r,$$

pues

$$\begin{aligned} \sum_{r=0}^{m} c_r (z+h)^r &= \sum_{r=0}^{m} c_r \sum_{k=0}^{m} (z^r)^{k)} = \sum_{k=0}^{m} \left(\sum_{r=0}^{m} c_r z^r \right)^{k)} \\ &= \sum_{k=0}^{m} f^{k)}(z) = f(z+h). \end{aligned}$$

Así mismo,

$$f(0+h) = \sum_{k=0}^{m} f^{k)}(0) = \sum_{r=0}^{m} c_r r! = f(h).$$

Sea

$$u_r(z) = r! \sum_{n=1}^{\infty} \frac{z^n}{(r+n)!}.$$

Teniendo en cuenta que

$$\frac{|z|^n}{\frac{(r+n)!}{r!}} < \frac{|z|^n}{n!},$$

es claro que la serie converge en todo $\mathbb{C}$ y que $|u_r(z)| < e^{|z|}$. Llamaremos

$$\epsilon_r(z) = \frac{u_r(z)}{e^{|z|}}.$$

Así, $|\epsilon_r(z)| < 1$.

Teorema 3.41 *Sea* $\phi(z) = \sum_{r=0}^{s} c_r z^r \in \mathbb{C}[z]$. *Sea* $\psi(z) = \sum_{r=0}^{s} c_r \epsilon_r(z) z^r$. *Entonces*

$$e^z \phi(h) = \phi(z+h) + \psi(z) e^{|z|}.$$

DEMOSTRACIÓN: En primer lugar

$$(z+h)^r = \sum_{k=0}^{r} \binom{r}{k} z^k h^{r-k} = \sum_{k=0}^{r} \frac{r!}{k!} z^k = r! \sum_{k=0}^{r} \frac{z^k}{k!},$$

y por otro lado

$$\begin{aligned} r!e^z - u_r(z)z^r &= r! \sum_{k=0}^{\infty} \frac{z^k}{k!} - r! \sum_{n=1}^{\infty} \frac{z^n}{(r+n)!} z^r = r! \sum_{k=0}^{\infty} \frac{z^k}{k!} - r! \sum_{k=r+1}^{\infty} \frac{z^k}{k!} \\ &= r! \sum_{k=0}^{r} \frac{z^k}{k!}, \end{aligned}$$

o sea, $(z+h)^r = r!e^z - u_r(z)z^r$, y por lo tanto

$$e^z h^r = (z+h)^r + u_r(z) z^r = (z+h)^r + e^{|z|} \epsilon_r(z) z^r.$$

Multiplicando por c_r y sumando en r obtenemos la igualdad buscada. ■

Teorema 3.42 *Sea* $m \geq 2$ *y* $f(x) \in \mathbb{Z}[x]$. *Definimos los polinomios* F_1 *y* F_2 *mediante:*

$$F_1(x) = \frac{x^{m-1}}{(m-1)!} f(x), \quad F_2(x) = \frac{x^m}{(m-1)!} f(x).$$

Entonces $F_1(h), F_2(h) \in \mathbb{Z}$, $F_1(h) \equiv f(0) \ (\text{mód } m)$ *y* $F_2(h) \equiv 0 \ (\text{mód } m)$.

DEMOSTRACIÓN: Sea $f(x) = \sum_{r=0}^{r} a_r x^r$, con $a_r \in \mathbb{Z}$. Entonces

$$F_1(x) = \sum_{r=0}^{s} a_r \frac{x^{r+m-1}}{(m-1)!}, \quad F_1(h) = \sum_{r=0}^{s} a_r \frac{(r+m-1)!}{(m-1)!} \in \mathbb{Z}$$

y m divide a cada sumando excepto quizá al primero, que es $a_0 = f(0)$. Por lo tanto $F_1(h) \equiv f(0) \ (\text{mód } m)$. Con F_2 se razona análogamente. ■

Como último preliminar recordemos que $p(x_1, \ldots, x_n) \in A[x_1, \ldots, x_n]$ es un polinomio *simétrico* si para toda permutación σ de las variables se cumple que $p(x_1, \ldots, x_n) = p\big(\sigma(x_1), \ldots, \sigma(x_n)\big)$. Los *polinomios simétricos elementales* de n variables son los polinomios $e_0, \ldots, e_n$ tales que e_i es la suma de todos los monomios posibles formados por i variables distintas. Por ejemplo, los polinomios simétricos elementales de tres variables son

$$e_0 = 1, \quad e_1 = x+y+z, \quad e_2 = xy + xz + yz, \quad e_3 = xyz.$$

Vamos a usar los dos resultados siguientes sobre polinomios elementales:

Todo polinomio simétrico $p(x_1,\ldots,x_n)$ *es de la forma* $q(e_1,\ldots,e_n)$, *para cierto polinomio* $q(x_1,\ldots,x_n)$.

Los coeficientes $(x-\alpha_1)\cdots(x-\alpha_n)$ *son* $(-1)^i e_i(\alpha_1,\ldots,\alpha_n)$, *para* $i=0,\ldots,n$.

Teorema 3.43 *El número* e *es trascendente.*

DEMOSTRACIÓN: Si e fuera algebraico sería la raíz de un polinomio con coeficientes enteros. Digamos $\sum\limits_{t=0}^{n} c_t e^t = 0$, con $n \geq 1$, $c_t \in \mathbb{Z}$, $c_0 \neq 0$ (si c_0 fuera 0 podríamos dividir entre e y quedarnos con un polinomio menor).

Sea p un primo tal que $p > n$ y $p > |c_0|$. Sea

$$\phi(x) = \frac{x^{p-1}}{(p-1)!}\big((x-1)(x-2)\cdots(x-n)\big)^p.$$

Por el teorema 3.41:

$$0 = \sum_{t=0}^{n} c_t e^t \phi(h) = \sum_{t=0}^{n} c_t \phi(t+h) + \sum_{t=0}^{n} c_t \psi(t) e^t = S_1 + S_2.$$

Tomando $m = p$, el teorema 3.42 nos da que $\phi(h) \in \mathbb{Z}$ y

$$\phi(h) \equiv (-1)^{pn}(n!)^p \text{ (mód } p).$$

Si $1 \leq t \leq n$, entonces

$$\phi(t+x) = \frac{(x+t)^{p-1}}{(p-1)!}\big((x+t-1)(x+t-2)\cdots x \cdots (x+t-n)\big)^p,$$

y sacando el factor x queda

$$\phi(t+h) = \frac{x^p}{(p-1)!} f(x),$$

con $f(x) \in \mathbb{Z}[x]$. Por el teorema 3.42 tenemos que $\phi(t+h) \in \mathbb{Z}$ y es un múltiplo de p. Ahora,

$$S_1 = \sum_{t=0}^{n} c_t \phi(t+h) \equiv c_0 \phi(h) \equiv c_0 (-1)^{pn}(n!)^p \not\equiv 0 \text{ (mód } p),$$

ya que $c_0 \neq 0$ y $p > n, |c_0|$. Así pues, $S_1 \in \mathbb{Z}$ y $S_1 \neq 0$, luego $|S_1| \geq 1$. Como $S_1 + S_2 = 0$, lo mismo vale para S_2, es decir, $|S_2| \geq 1$. Por otro lado, sea $\phi(x) = \sum\limits_{r=0}^{s} a_r x^r$. Entonces

$$|\psi(t)| = \left|\sum_{r=0}^{s} a_r \epsilon_r(t) t^r\right| \leq \sum_{r=0}^{s} |a_r|\,|\epsilon_r(t)| t^r \leq \sum_{r=0}^{s} |a_r| t^r.$$

Observemos que $|a_r|$ es el coeficiente de grado r del polinomio

$$\frac{x^{p-1}}{(p-1)!}\big((x+1)\cdots(x+n)\big)^p.$$

Basta ver que si b_i es el coeficiente de grado i de $\big((x-1)\cdots(x-n)\big)^p$, entonces $|b_i|$ es el coeficiente de grado i de $\big((x+1)\cdots(x+n)\big)^p$, pero

$$\begin{aligned}|b_i| &= |(-1)^{pn-i}e_{np-i}(1,\ldots,n,\ldots,1,\ldots,n)|\\ &= (-1)^{pn-i}e_{np-i}(-1,\ldots,-n,\ldots,-1,\ldots,-n).\end{aligned}$$

Por lo tanto podemos concluir:

$$|\psi(t)| \leq \sum_{r=0}^{s}|a_r|t^r = \frac{t^{p-1}}{(p-1)!}\big((t+1)(t+2)\cdots(t+n)\big)^p,$$

y en definitiva:

$$|\psi(t)| \leq (t+1)(t+2)\cdots(t+n)\frac{\big(t(t+1)\cdots(t+n)\big)^{p-1}}{(p-1)!}.$$

Pero esta expresión tiende a 0 cuando p tiende a infinito (por la convergencia de la serie de la función exponencial). En consecuencia, tomando p suficientemente grande podemos exigir que $S_2 = \sum_{t=0}^{n} c_t\psi(t)e^t$ cumpla $|S_2| < 1$, cuando hemos demostrado lo contrario para todo p. Esto prueba que e es trascendente.

■

La trascendencia de π es algo más complicada de probar. Además de los teoremas 3.41 y 3.42 necesitaremos otro resultado auxiliar:

Teorema 3.44 *Sea $p(x) = dx^m + d_1x^{m-1} + \cdots + d_{m-1}x + d_m \in \mathbb{Z}[x]$, sean $\alpha_1,\ldots,\alpha_m$ sus raíces en $\mathbb{C}$ y sea $q(x_1,\ldots,x_m) \in \mathbb{Z}[x_1,\ldots,x_m]$ un polinomio simétrico. Entonces $q(d\alpha_1,\ldots,d\alpha_m) \in \mathbb{Z}$.*

Demostración: Claramente

$$d^{m-1}p(x) = (dx)^m + d_1(dx)^{m-1} + dd_2(dx)^{m-2} + \cdots + d^{m-2}d_{m-1}(dx) + d^{m-1}d_m.$$

O sea, $d^{m-1}p(x) = r(dx)$, donde

$$r(x) = x^m + d_1x^{m-1} + dd_2x^{m-2} + \cdots + d^{m-2}d_{m-1}x + d^{m-1}d_m \in \mathbb{Z}[x]$$

es un polinomio mónico y sus raíces son obviamente $d\alpha_1,\ldots,d\alpha_m$. Consecuentemente, $r(x) = (x-d\alpha_1)\cdots(x-d\alpha_m)$ y los coeficientes de $r(x)$ son $(-1)^ie_i(d\alpha_1,\ldots,d\alpha_m)$ para $i = 0,\ldots m$. Así pues, $e_i(d\alpha_1,\ldots,d\alpha_m) \in \mathbb{Z}$ para $i = 1,\ldots,m$.

Por otro lado sabemos que $q(x_1,\ldots,x_m) = r(e_1,\ldots,e_m)$ para cierto polinomio $r(x_1,\ldots,x_m) \in \mathbb{Z}[x_1,\ldots,x_m]$, luego

$$q(d\alpha_1,\ldots,d\alpha_m) = r\big(e_1(d\alpha_1,\ldots,d\alpha_m),\ldots,e_m(d\alpha_1,\ldots,d\alpha_m)\big) \in \mathbb{Z}.$$

■

Teorema 3.45 *El número π es trascendente.*

DEMOSTRACIÓN: Si π es algebraico también lo es $i\pi$. Sea

$$dx^m + d_1x^{m-1} + \cdots + d_{m-1}x + d_m \in \mathbb{Z}[x]$$

un polinomio tal que $d \neq 0$ y con raíz $i\pi$. Sean $\omega_1, \ldots, \omega_m$ sus raíces en $\mathbb{C}$. Como una de ellas es $i\pi$ y $e^{i\pi} + 1 = 0$, tenemos que $(1 + e^{\omega_1}) \cdots (1 + e^{\omega_m}) = 0$, o sea,

$$1 + \sum_{t=1}^{2^m-1} e^{\alpha_t} = 0,$$

donde $\alpha_1, \ldots, \alpha_{2^m-1}$ son $\omega_1, \ldots, \omega_m, \omega_1+\omega_2, \ldots, \omega_{m-1}+\omega_m, \ldots, \omega_1+\cdots+\omega_m$.

Supongamos que $c-1$ de ellos son nulos y $n = 2^m - 1 - (c-1)$ no son nulos. Ordenémoslos como $\alpha_1, \ldots, \alpha_n, 0, \ldots, 0$. Así $c \geq 1$ y

$$c + \sum_{t=1}^{n} e^{\alpha_t} = 0. \tag{3.7}$$

Notemos lo siguiente:

$$e_i(x_1, \ldots, x_n) = e_i(x_1, \ldots, x_n, 0, \ldots, 0),$$

donde e_i es a la izquierda el polinomio simétrico elemental de n variables y a la derecha el de $2^m - 1$ variables. Por lo tanto

$$e_i(d\alpha_1, \ldots, d\alpha_n) = e_i(d\alpha_1, \ldots, d\alpha_{2^m-1})$$

$$= e_i(d\omega_1, \ldots, d\omega_m, d\omega_1 + d\omega_2, \ldots, d\omega_{m-1} + d\omega_m, \ldots, d\omega_1 + \cdots + d\omega_m).$$

Sea $q_i(x_1, \ldots, x_m) = e_i(x_1, \ldots, x_m, x_1+x_2, \ldots, x_{m-1}+x_m, \ldots, x_1+\cdots+x_m)$.

Claramente se trata de polinomios simétricos con coeficientes enteros y

$$e_i(d\alpha_1, \ldots, d\alpha_n) = q_i(d\omega_1, \ldots, d\omega_m),$$

luego el teorema anterior nos permite afirmar que $e_i(d\alpha_1, \ldots, d\alpha_n) \in \mathbb{Z}$.

Si $s(x_1, \ldots, x_n) \in \mathbb{Z}[x_1, \ldots, x_n]$ es simétrico, entonces s depende polinómicamente de los e_i, luego $s(d\alpha_1, \ldots, d\alpha_n) \in \mathbb{Z}$. Sea p un primo tal que $p > |d|$, $p > c$, $p > |d^n\alpha_1 \cdots \alpha_n|$. Sea

$$\phi(x) = \frac{d^{np+p-1}x^{p-1}}{(p-1)!}\big((x-\alpha_1)\cdots(x-\alpha_n)\big)^p.$$

Multiplicamos (3.7) por $\phi(h)$ y aplicamos el teorema 3.41. Nos queda

$$c\phi(h) + \sum_{t=1}^{n} \phi(\alpha_t + h) + \sum_{t=1}^{n} \psi(\alpha_t)e^{|\alpha_t|} = S_0 + S_1 + S_2 = 0.$$

Ahora,

$$\phi(x) = \frac{x^{p-1}}{(p-1)!}d^{p-1}\big((dx - d\alpha_1)\cdots(dx - d\alpha_n)\big)^p.$$

Los coeficientes de $(y - d\alpha_1) \cdots (y - d\alpha_n)$ son polinomios simétricos elementales sobre $d\alpha_1, \ldots, d\alpha_n$, luego son enteros, según hemos visto antes. De aquí se sigue que también son enteros los coeficientes de $(dx - d\alpha_1) \cdots (dx - d\alpha_n)$, con lo que

$$\phi(x) = \frac{x^{p-1}}{(p-1)!} \sum_{r=0}^{np} g_r x^r, \quad \text{donde cada } g_r \in \mathbb{Z}.$$

Por el teorema 3.42 tenemos que $\phi(h) \in \mathbb{Z}$ y $\phi(h) \equiv g_0 \pmod{p}$. Concretamente, $g_0 = (-1)^{pn} d^{p-1} (d\alpha_1 \cdots d\alpha_n)^p$, luego por la elección de p resulta que $p \nmid g_0$ (aquí es importante que $d\alpha_1 \cdots d\alpha_n \in \mathbb{Z}$ porque es el término independiente de $(y - d\alpha_1) \cdots (y - d\alpha_n)$. Como $p \nmid c$, resulta que $p \nmid S_0 = c\phi(h)$.

Nos ocupamos ahora de S_1. Tenemos que

$$\phi(\alpha_t + x) = \frac{d^{np+p-1}(\alpha_t + x)^{p-1}}{(p-1)!} \big((x + \alpha_t - \alpha_1) \cdots (x + \alpha_t - \alpha_{t-1})$$

$$x(x + \alpha_t - \alpha_{t+1}) \cdots (x + \alpha_t - \alpha_n)\big)^p$$

$$= \frac{x^p}{(p-1)!} d^p (d\alpha_t + dx)^{p-1} \big((dx + d\alpha_t - d\alpha_1) \cdots (dx + d\alpha_t - d\alpha_{t-1})$$

$$(dx + d\alpha_t - d\alpha_{t+1}) \cdots (dx + d\alpha_t - d\alpha_n)\big)^p$$

$$= \frac{x^p}{(p-1)!} \sum_{r=0}^{np-1} f_{rt} x^r,$$

donde $f_{rt} = f_r(d\alpha_t, d\alpha_1, \ldots, d\alpha_{t-1}, d\alpha_{t+1}, \ldots, d\alpha_n)$ y f_r es un polinomio simétrico respecto a todas las variables excepto la primera, con coeficientes enteros y que no depende de t. En efecto, consideramos el polinomio

$$\big(y - (-x_1)\big)^{p-1} \Big(\big(y - (x_2 - x_1)\big) \cdots \big((y - (x_n - x_1)\big)\Big)^p.$$

Sus coeficientes son los polinomios simétricos elementales actuando sobre $-x_1$ ($p-1$ veces) y sobre $x_2 - x_1, \ldots, x_n - x_1$ (p veces cada uno), luego son polinomios simétricos en $x_2, \ldots, x_n$. Digamos que el polinomio es $\sum\limits_{r=0}^{np-1} s_r(x_1, \ldots, x_n) y^r$. Entonces

$$\phi(\alpha_t + x) = \frac{x^p}{(p-1)!} \sum_{r=0}^{np-1} d^p s_r(d\alpha_t, d\alpha_1, \ldots, d\alpha_{t-1}, d\alpha_{t+1}, \ldots, d\alpha_n) d^r x^r,$$

es decir, $f_r = d^{r+p} s_r(x_1, \ldots, x_n)$. Por lo tanto,

$$\sum_{t=1}^{n} \phi(\alpha_t + x) = \frac{x^p}{(p-1)!} \sum_{r=0}^{np-1} \left(\sum_{t=1}^{n} f_{rt} \right) x^r,$$

pero

$$\sum_{t=1}^{n} f_{rt} = \sum_{t=1}^{n} f_r(d\alpha_t, d\alpha_1, \ldots, d\alpha_{t-1}, d\alpha_{t+1}, \ldots, d\alpha_n),$$

y el polinomio $\sum\limits_{t=1}^{n} f_r(x_t, x_1, \ldots, x_{t-1}, x_{t+1}, \ldots x_n)$ es simétrico (respecto a todas las variables), luego $F_r = \sum\limits_{t=1}^{n} f_{rt}$ depende simétricamente de $d\alpha_1, \ldots, d\alpha_n$ y por lo tanto es un entero. Así pues,

$$\sum_{t=1}^{n} \phi(\alpha_t + x) = \frac{x^p}{(p-1)!} \sum_{r=0}^{np-1} F_r x^r,$$

y por el teorema 3.42 concluimos que $S_1 = \sum\limits_{t=1}^{n} \phi(\alpha_t + h) \in \mathbb{Z}$ y es múltiplo de p (aquí hemos usado que $(\phi_1 + \phi_2)(h) = \phi_1(h) + \phi_2(h)$, lo cual es evidente).

Esto nos da que $S_0 + S_1 \in \mathbb{Z}$ y no es un múltiplo de p. En particular $|S_0 + S_1| \geq 1$ y, por la ecuación $S_0 + S_1 + S_2 = 0$ resulta que también $S_2 \in \mathbb{Z}$ y $|S_2| \geq 1$. Como en el caso de e, ahora probaremos lo contrario.

Sea $\psi(x) = \sum\limits_{r=0}^{np+p-1} c_r \epsilon_r(x) x^r$. Entonces

$$\begin{aligned} |\psi(x)| &= \left| \sum_{r=0}^{np+p-1} c_r \epsilon_r(x) x^r \right| \leq \sum_{r=0}^{np+p-1} |c_r|\,|x^r| \\ &\leq \frac{|d|^{np+p-1}|x|^{p-1}}{(p-1)!} \big((|x| + |\alpha_1|) \cdots (|x| + |\alpha_n|)\big)^p \end{aligned}$$

(por el mismo razonamiento que en la prueba de la trascendencia de e) y así

$$|\psi(x)| \leq \frac{M^{2np+2p-2}}{(p-1)!},$$

donde M es una cota que no depende de p.

Como $M^{2np+2p-2} \leq M^{2np+2p} = M^{(2n+2)p} = K^p = KK^{p-1}$, tenemos que

$$|\psi(x)| \leq K\, \frac{K^{p-1}}{(p-1)!}$$

y la sucesión converge a 0 cuanto p tiende a infinito, pues la serie converge a Ke^K. Esto para cada x fijo. Tomando un primo p suficientemente grande podemos exigir que $|S_2| \leq \sum\limits_{t=1}^{n} |\psi(\alpha_t)| e^{|\alpha_t|} < 1$, con lo que llegamos a una contradicción. ■

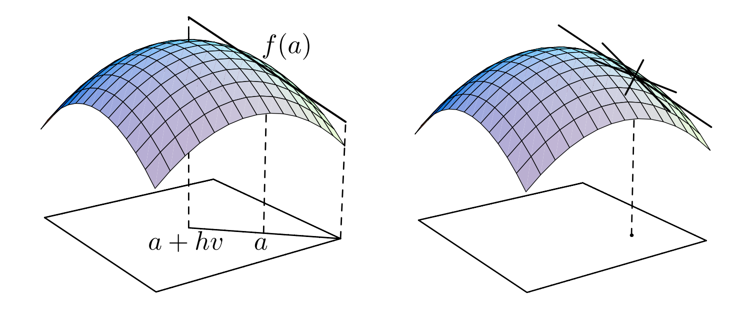
Capítulo IV

Cálculo diferencial de varias variables

Este capítulo está dedicado a generalizar a funciones de varias variables las ideas que introdujimos en el capítulo anterior. Básicamente se trata de estudiar cómo varía una función de varias variables cuando incrementamos éstas infinitesimalmente. Más concretamente estudiaremos una función $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$, donde A es un abierto, aunque a efectos de interpretar la teoría nos centraremos de momento en el caso en que $m = 1$.

4.1 Diferenciación

Pensemos por ejemplo en una función $f : \mathbb{R}^2 \longrightarrow \mathbb{R}$. Mientras una función de una variable derivable en un punto tiene asociada una única recta tangente que la aproxima, una función de dos variables tiene (o puede tener) una tangente distinta para cada dirección. La figura muestra (a la izquierda) una tangente a la gráfica de f en un punto, y a la derecha vemos varias tangentes distintas en ese mismo punto.



Intuitivamente está claro qué es la recta tangente a una superficie en un punto y en una dirección. Ahora vamos a caracterizar matemáticamente este

concepto. Tenemos un punto $a \in \mathbb{R}^n$ y una recta que pasa por a. Ésta queda determinada por un vector $v \in \mathbb{R}^n$ no nulo. Podemos suponer $\|v\| = 1$ (mientras no se indique lo contrario, todas las normas que consideraremos serán euclídeas). Los puntos de la recta son los de la forma $a+hv$, con $h \in \mathbb{R}$. Más concretamente, el punto $a+hv$ es el que se encuentra a una distancia $|h|$ de a sobre dicha recta (el signo de h distingue los dos puntos en estas condiciones).

Buscamos una recta que se parece a la gráfica de f alrededor del punto a. Si la gráfica fuera rectilínea en la dirección considerada, su pendiente vendría dada por

$$\frac{f(a+hv)-f(a)}{h},$$

para cualquier $h \neq 0$. Si no es así, entonces esta expresión se parecerá más a la pendiente que buscamos cuanto menor sea h. Ello nos lleva a la definición siguiente:

Definición 4.1 Dada una función $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ definida en un abierto, un punto $a \in A$ y un vector $v \in \mathbb{R}^n$ no nulo, llamaremos *derivada direccional* de f en a y en la dirección de v al vector

$$f'(a;v) = \lim_{h\to 0} \frac{f(a+hv)-f(a)}{h} \in \mathbb{R}^m.$$

Es fácil ver que si existe $f'(a;v)$ entonces existe $f'(a;\lambda v) = \lambda f'(a,v)$ para todo $\lambda \in \mathbb{R}\setminus\{0\}$. Por lo tanto no perdemos generalidad si suponemos $\|v\| = 1$. Si existe $f'(a;v)$, para valores pequeños de h tenemos la aproximación

$$f(a+hv) \approx f(a) + h\,f'(a;v),$$

con lo que la expresión $h\,f'(a;v)$ aproxima el incremento que experimenta $f(a)$ cuando la variable se incrementa h unidades en la dirección de v.

En el caso $m = 1$ la función $a+hv \mapsto f(a)+h\,f'(a;v)$ se llama *recta tangente* a la gráfica de f en a. Es claro que se trata de la recta que pretendíamos caracterizar.

Como en el caso de una variable, si una función $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ admite derivada direccional en la dirección de v y en todo punto de A, entonces tenemos definida una función

$$f'(\ ;v) : A \longrightarrow \mathbb{R}^m.$$

Respecto al cálculo de derivadas direccionales, las propiedades de los límites nos dan en primer lugar que si $f(x) = \big(f_1(x),\ldots,f_m(x)\big)$, entonces

$$f'(a;v) = \big(f_1'(a;v),\ldots,f_m'(a;v)\big),$$

entendiendo que la derivada de f existe si y sólo si existen las derivadas de todas las funciones coordenadas f_i. Por consiguiente el cálculo de derivadas direccionales se reduce al caso de funciones $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}$. A su vez éstas se reducen al cálculo de derivadas de funciones de una variable. Efectivamente, basta considerar la función $\phi(h) = f(a+hv)$. El hecho que que A sea abierto implica claramente que ϕ está definida en un entorno de 0 y comparando las definiciones es claro que $f'(a;v) = \phi'(0)$.

Ejemplo Vamos a calcular la derivada de $f(x,y) = x^2y^2$ en el punto $(2,1)$ y en la dirección $(-1,1)$. Para ello consideramos

$$\phi(h) = f(2-h, 1+h) = (2-h)^2(1+h)^2.$$

Entonces $\phi'(h) = -2(2-h)(1+h)^2 + 2(2-h)^2(1+h)$ y $\phi'(0) = 4$. ■

Existen unas derivadas direccionales especialmente simples de calcular y especialmente importantes en la teoría. Se trata de las derivadas en las direcciones de la base canónica de $\mathbb{R}^n$.

Definición 4.2 Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ una función definida en un abierto y $a \in A$. Se define la *derivada parcial* de f respecto a la i-ésima variable en el punto a como

$$D_i f(a) = \frac{\partial f}{\partial x_i}(a) = f'(a; e_i),$$

donde $e_i = (0, \ldots, 1, \ldots, 0)$ es el vector con un 1 en la posición i-ésima.

Explícitamente:

$$D_i f(a) = \lim_{h \to 0} \frac{f(a_1, \ldots, a_i + h, \ldots, a_n) - f(a)}{h},$$

luego si h es pequeño

$$f(a_1, \ldots, a_i + h, \ldots, a_n) \approx f(a) + h\, D_i f(a).$$

En otras palabras, la expresión $h\, D_i f(a)$ aproxima el incremento que experimenta $f(a)$ cuando incrementamos h unidades la variable x_i.

Si f admite derivada parcial i-ésima en todos los puntos de A entonces tenemos definida la función $D_i f : A \longrightarrow \mathbb{R}^m$.

El cálculo de las derivadas parciales de una función $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ es mucho más simple que el de las derivadas direccionales en general, pues podemos considerar la función $\phi(x_i) = f(a_1, \ldots, x_i, \ldots, a_n)$ y entonces es claro que $D_i f(a) = \phi'(a_i)$. Notar que si f es una función de una variable, entonces $\phi = f$, luego la derivada parcial respecto de la única variable coincide con la derivada de f en el sentido del capítulo anterior. Un poco más en general, si $f : A \subset \mathbb{R} \longrightarrow \mathbb{R}^m$ llamaremos también *derivada* de f a su única derivada parcial en un punto a, y la representaremos también por $f'(a)$.

Ejemplo Las derivadas parciales de la función $f(x,y) = x^2y^3$ son

$$\frac{\partial f}{\partial x} = 2xy^3, \quad \frac{\partial f}{\partial y} = 3x^2y^2.$$

En efecto, para calcular la parcial respecto de x en un punto (x_0, y_0) hay que derivar la función $x \mapsto x^2y_0^3$ en x_0. La derivada es obviamente $2x_0y_0^3$. En la práctica podemos ahorrarnos los subíndices: para derivar x^2y^3 respecto de x

basta considerar a y como una constante (la segunda coordenada del punto en que derivamos) y derivar respecto de x. Lo mismo vale para y. ■

Es claro que todas las reglas de derivación de funciones de una variable pueden ser usadas en el cálculo de derivadas parciales.

Ejercicio: Sean dos funciones derivables $f, g : I \subset \mathbb{R} \longrightarrow \mathbb{R}^n$. Probar la regla de derivación $(fg)' = f'g + fg'$. Si $n = 3$ se cumple también $(f \wedge g)' = f' \wedge g + f \wedge g'$.

El hecho de que una función admita derivadas direccionales en un punto no puede ser equiparado a la derivabilidad de una función de una variable. Por ejemplo, la existencia de derivadas direccionales no implica siquiera la continuidad de la función en el punto. La generalización adecuada del concepto de función derivable es el concepto de "función diferenciable", que vamos a introducir ahora.

Recordemos que si una función de una variable f tiene derivada en un punto a, entonces podemos definir su diferencial en a, que es una aplicación lineal $df(a) : \mathbb{R} \longrightarrow \mathbb{R}$ con la propiedad de que $f(a) + df(a)(x - a)$ es una recta que "se confunde" con f alrededor de a. Una función de varias variables será diferenciable cuando exista una aplicación lineal que juegue un papel análogo:

Definición 4.3 Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ una función definida en un abierto A. Sea $a \in A$. Diremos que f es *diferenciable* en A si existe una aplicación lineal $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ tal que

$$\lim_{v \to 0} \frac{f(a + v) - f(a) - \phi(v)}{\|v\|} = 0.$$

Para analizar esta definición conviene comenzar probando lo siguiente:

Teorema 4.4 *Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ una función diferenciable en un punto $a \in A$. Sea $\phi : \mathbb{R}^n \longrightarrow \mathbb{R}^m$ una aplicación lineal que cumpla la definición anterior. Entonces, para cada vector $v \in \mathbb{R}^n$ no nulo existe $f'(a; v)$ y además $\phi(v) = f'(a; v)$.*

DEMOSTRACIÓN: Obviamente hv tiende a 0 cuando h tiende a 0. Por lo tanto, restringiendo el límite de la definición de diferenciabildad concluimos que

$$\lim_{h \to 0} \frac{f(a + hv) - f(a) - \phi(hv)}{\|hv\|} = 0.$$

Usando que ϕ y la norma son lineales vemos que

$$\lim_{h \to 0} \frac{1}{\|v\|} \left(\frac{f(a + hv) - f(a)}{|h|} - \frac{h}{|h|}\phi(v) \right) = 0.$$

Claramente podemos eliminar el factor $1/\|v\|$ sin que el límite varíe. Ahora multiplicamos por la función $\mathbb{R} \setminus \{0\} \longrightarrow \{\pm 1\}$ dada por $h \mapsto |h|/h$ y usamos que el producto de una función que tiende a 0 por otra acotada tiende a 0:

$$\lim_{h \to 0} \frac{f(a + hv) - f(a)}{h} - \phi(v) = 0,$$

Por lo tanto existe

$$f'(a;v) = \lim_{h\to 0} \frac{f(a+hv)-f(a)}{h} = \phi(v).$$

■

En particular vemos que si f es diferenciable en a existe una única aplicación lineal ϕ que cumple la definición de diferenciabilidad, a saber, la dada por $\phi(v) = f'(a;v)$.

Definición 4.5 Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ una función diferenciable en un punto $a \in A$. Llamaremos *diferencial* de f en a a la única aplicación lineal, representada por $df(a) : \mathbb{R}^n \longrightarrow \mathbb{R}^m$, que cumple

$$\lim_{v\to 0} \frac{f(a+v)-f(a)-df(a)(v)}{\|v\|} = 0.$$

El teorema anterior afirma que para todo $v \in \mathbb{R}^n \setminus \{0\}$ se cumple

$$df(a)(v) = f'(a;v).$$

Consideremos el caso $m = 1$. Entonces, para un vector unitario v, la función

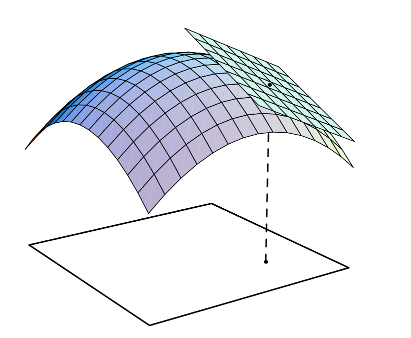
$$a + hv \mapsto f(a) + df(a)(hv) = f(a) + h\,df(a)(v) = f(a) + hf'(a;v)$$

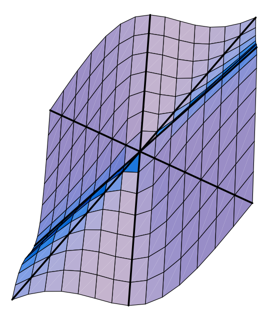
es la recta tangente a f por a y en la dirección de v. Cuando h varía en $\mathbb{R}$ y v varía entre los vectores unitarios, el punto $x = a + hv$ varía en todo $\mathbb{R}^n$ y la aplicación $x \mapsto f(a) + df(a)(x-a)$ recorre todos los puntos de todas las rectas tangentes a f por a. Puesto que se trata de una aplicación afín, dichas tangentes forman un hiperplano.

En resumen, hemos probado que para que una aplicación (con $m = 1$) sea diferenciable en un punto a es necesario que tenga rectas tangentes por a en todas las direcciones y que además éstas formen un hiperplano. A este hiperplano se le llama *hiperplano tangente* a la gráfica de f en a. De la propia definición de diferencial (haciendo $v = x - a$) se sigue que para puntos x cercanos a a se cumple

$$f(x) \approx f(a) + df(a)(x-a).$$

El miembro derecho es precisamente el hiperplano tangente a f en a. Esta expresión indica, pues, que dicho hiperplano aproxima a f alrededor de a.





La figura de la derecha muestra la gráfica de la función

$$f(x,y) = \begin{cases} \frac{y^3-x^3}{x^2+y^2} & \text{si } (x,y) \neq (0,0) \\ 0 & \text{si } (x,y) = (0,0) \end{cases}$$

Vemos también cuatro tangentes en $(0,0)$. Como no se encuentran sobre el mismo plano, concluimos que esta función no es diferenciable en $(0,0)$.

Pasamos ahora al cálculo de la diferencial de una función. Como primeras observaciones elementales notamos que si f es lineal entonces $df(a) = f$ y si f es constante (alrededor de a) entonces $df(a) = 0$ (la aplicación nula). Ambos hechos se demuestran comprobando que con las elecciones indicadas para ϕ se cumple trivialmente la definición de diferencial.

Para una función $f(x) = \big(f_1(x), \ldots, f_m(x)\big)$, las propiedades de los límites nos dan que f es diferenciable en un punto a si y sólo si lo es cada función coordenada f_i, y en tal caso

$$df(a)(v) = \big(df_1(a)(v), \ldots, df_m(a)(v)\big).$$

Para determinar $df(a)$ es suficiente conocer su matriz en las bases canónicas de $\mathbb{R}^n$ y $\mathbb{R}^m$. Dicha matriz tiene por filas las imágenes de los vectores e_i de la base canónica.

Definición 4.6 Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ una función diferenciable en un punto $a \in A$. Llamaremos *matriz jacobiana* de f en a a la matriz $Jf(a)$ que tiene por filas a las derivadas parciales $D_i f(a)$.

Más concretamente, si $f(x) = \big(f_1(x), \ldots, f_m(x)\big)$, el coeficiente de la fila i, columna j de $Jf(a)$ es $D_i f_j(a)$.

Si $m = 1$, se llama *vector gradiente* de f en a al vector formado por las derivadas parciales de f en a. Se representa:

$$\nabla f(a) = \big(D_1 f(a), \ldots, D_n f(a)\big).$$

Si e_i es el vector i-ésimo de la base canónica, sabemos que

$$df(a)(e_i) = f'(a; e_i) = D_i f(a),$$

luego la matriz jacobiana de f es simplemente la matriz de $df(a)$ en las bases canónicas de $\mathbb{R}^n$ y $\mathbb{R}^m$. Así pues,

$$df(a)(v) = vJf(a).$$

Cuando $m = 1$, usando el producto escalar en lugar del producto de matrices tenemos también

$$df(a)(v) = \nabla f(a)v = \frac{\partial f}{\partial x_1}(a)\, v_1 + \cdots + \frac{\partial f}{\partial x_n}(a)\, v_n.$$

Consideremos en particular la función polinómica x_i, es decir, la función $\mathbb{R}^n \longrightarrow \mathbb{R}$ dada por $(x_1, \ldots, x_n) \mapsto x_i$. Es claro que $\nabla x_i(a) = e_i$, luego $dx_i(a)(v) = v_i$. Por consiguiente, la ecuación anterior puede escribirse como

$$df(a)(v) = \frac{\partial f}{\partial x_1}(a)\, dx_1(a)(v) + \cdots + \frac{\partial f}{\partial x_n}(a)\, dx_n(a)(v).$$

Como esto es válido para todo v, tenemos la ecuación funcional

$$df(a) = \frac{\partial f}{\partial x_1}(a)\, dx_1(a) + \cdots + \frac{\partial f}{\partial x_n}(a)\, dx_n(a).$$

Si $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ es diferenciable en todos los puntos de A podemos considerar df, dx_i como funciones de A en el espacio de aplicaciones lineales de $\mathbb{R}^n$ en $\mathbb{R}$ y la ecuación anterior nos da

$$df = \frac{\partial f}{\partial x_1}\, dx_1 + \cdots + \frac{\partial f}{\partial x_n}\, dx_n.$$

Esta fórmula expresa que si cada variable experimenta un incremento infinitesimal dx_i entonces la función f experimenta un incremento df de la forma que se indica. Como en el caso de una variable, la expresión ha de entenderse en realidad como una igualdad funcional que a cada vector de incrementos $(\Delta x_1, \ldots, \Delta x_n)$ le asigna una aproximación del incremento Δf que experimenta la función.

Similarmente, en el caso $m > 1$ tenemos

$$df = (df_1, \ldots, df_m) = (dx_1, \ldots, dx_n)\, Jf.$$

Ejemplo Consideremos la función $]0, +\infty[\times]-\pi, \pi[\longrightarrow \mathbb{R}^2$ dada por

$$\begin{aligned} x &= \rho \cos \theta, \\ y &= \rho \operatorname{sen} \theta. \end{aligned}$$

Podríamos demostrar que es diferenciable aplicando la definición, pero más adelante será inmediato (teorema 4.11), así que vamos a aceptar que lo es y calcularemos su diferencial. Para ello calculamos la matriz jacobiana:

$$J(x, y)(\rho, \theta) = \begin{pmatrix} \dfrac{\partial x}{\partial \rho} & \dfrac{\partial y}{\partial \rho} \\ \dfrac{\partial x}{\partial \theta} & \dfrac{\partial y}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos\theta & \operatorname{sen}\theta \\ -\rho \operatorname{sen}\theta & \rho\cos\theta \end{pmatrix}$$

Por consiguiente

$$\begin{aligned} (dx, dy) &= (d\rho, d\theta) \begin{pmatrix} \cos\theta & \operatorname{sen}\theta \\ -\rho \operatorname{sen}\theta & \rho\cos\theta \end{pmatrix} \\ &= (\cos\theta\, d\rho - \rho \operatorname{sen}\theta\, d\theta, \operatorname{sen}\theta\, d\rho + \rho\cos\theta\, d\theta), \end{aligned}$$

o más claramente:

$$\begin{aligned} dx &= \cos\theta\, d\rho - \rho \operatorname{sen}\theta\, d\theta, \\ dy &= \operatorname{sen}\theta\, d\rho + \rho\cos\theta\, d\theta. \end{aligned}$$

También podríamos haber calculado dx y dy de forma independiente. ■

4.2 Propiedades de las funciones diferenciables

Vamos a estudiar las funciones diferenciables. Entre otras cosas obtendremos un criterio sencillo que justificará la diferenciabilidad de la mayoría de funciones de interés. Comenzamos observando que en funciones de una variable la diferenciabilidad equivale a la derivabilidad.

Teorema 4.7 *Sea $f : A \subset \mathbb{R} \longrightarrow \mathbb{R}^m$ una función definida en un abierto A y sea $a \in A$. Entonces f es diferenciable en a si y sólo si existe la derivada de f en a. Además en tal caso $df(a)(h) = f'(a)h$.*

Demostración: Si f es derivable en a entonces existe

$$f'(a) = \lim_{h\to 0} \frac{f(a+h) - f(a)}{h} = k,$$

luego

$$\lim_{h\to 0} \frac{f(a+h) - f(a) - kh}{h} = 0,$$

y si multiplicamos por la función acotada $h/|h|$ el límite sigue siendo 0, es decir, tenemos

$$\lim_{h\to 0} \frac{f(a+h) - f(a) - kh}{|h|} = 0,$$

lo que indica que f es diferenciable y que $df(a)(h) = f'(a)h$.

El recíproco se prueba igualmente, partiendo de que $df(a)(h) = kh$ se llega a que existe $f'(a) = k$. ■

Teorema 4.8 *Si $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ es diferenciable en un punto a, entonces f es continua en a.*

Demostración: Tenemos que

$$\lim_{v\to 0} \frac{f(a+v) - f(a) - df(a)(v)}{\|v\|} = 0.$$

Multiplicamos por $\|v\|$, que también tiende a 0, con lo que

$$\lim_{v\to 0} f(a+v) - f(a) - df(a)(v) = 0.$$

La aplicación $df(a)$ es lineal, luego es continua, luego tiende a 0, luego

$$\lim_{v\to 0} f(a+v) - f(a) = 0,$$

y esto equivale a

$$\lim_{x\to a} f(x) = f(a),$$

luego f es continua en a. ■

Las propiedades algebraicas de la derivabilidad de funciones son válidas también para la diferenciabilidad:

Teorema 4.9 *Sean f y g funciones diferenciables en un punto a. Entonces*

a) $f+g$ es diferenciable en a y $d(f+g)(a) = df(a) + dg(a)$.

b) Si $\alpha \in \mathbb{R}$ entonces αf es diferenciable en a y $d(\alpha f)(a) = \alpha\, df(a)$.

c) fg es diferenciable en a y $d(fg)(a) = g(a)df(a) + f(a)dg(a)$.

d) si $g(a) \neq 0$ entonces f/g es diferenciable en a y

$$d(f/g)(a) = \frac{g(a)df(a) - f(a)dg(a)}{g^2(a)}.$$

Demostración: Veamos por ejemplo la propiedad c). Llamemos

$$E(v) = \frac{f(a+v) - f(a) - df(a)(v)}{\|v\|}, \quad F(v) = \frac{g(a+v) - g(a) - dg(a)(v)}{\|v\|}.$$

Ambas funciones están definidas en un entorno de 0 y tienden a 0. Además

$$f(a+v) - f(a) = df(a)(v) + \|v\|E(v), \quad g(a+v) - g(a) = dg(a)(v) + \|v\|F(v).$$

Entonces

$$(fg)(a+v) - (fg)(a) = f(a+v)g(a+v) - f(a)g(a+v) + f(a)g(a+v) - f(a)g(a)$$

$$= \big(f(a+v) - f(a)\big)g(a+v) + f(a)\big(g(a+v) - g(a)\big).$$

Sustituimos $f(a+v) - f(a)$, $g(a+v)$ y $g(a+v) - g(a)$ usando las igualdades anteriores. Al operar queda

$$(fg)(a+v) - (fg)(a) - \big(g(a)df(a) + f(a)dg(a)\big) = df(a)(v)dg(a)(v)$$

$$+\|v\|\big(df(a)(v)F(v) + E(v)g(a) + E(v)dg(a)(v) + f(a)F(v)\big) + \|v\|^2E(v)F(v).$$

Hay que probar que el miembro derecho dividido entre $\|v\|$ tiende a 0. El único término para el que esto no es inmediato es

$$\frac{df(a)(v)dg(a)(v)}{\|v\|},$$

pero $\|df(a)(v)dg(a)(v)\| \leq \|df(a)\|\,\|dg(a)\|\,\|v\|^2$, luego la norma del cociente está mayorada por

$$\|df(a)\|\,\|dg(a)\|\,\|v\|,$$

que tiende a 0. ■

Veamos ahora la versión en varias variables de la regla de la cadena.

Teorema 4.10 (Regla de la cadena) *Consideremos $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ y $g : B \subset \mathbb{R}^m \longrightarrow \mathbb{R}^k$ de modo que $f[A] \subset B$. Si f es diferenciable en un punto $a \in A$ y g es diferenciable en $f(a)$, entonces $f \circ g$ es diferenciable en a y*

$$d(f \circ g)(a) = df(a) \circ dg\big(f(a)\big).$$

DEMOSTRACIÓN: Llamemos $h = f \circ g$ y $b = f(a)$. Dado un $v \in \mathbb{R}^n$ tal que $a + v \in A$, tenemos

$$h(a+v) - h(a) = g(f(a+v)) - g(f(a)) = g(b+u) - g(b),$$

donde $u = f(a+v) - f(a)$. Consideremos las funciones

$$E(v) = \frac{f(a+v) - f(a) - df(a)(v)}{\|v\|}, \quad F(u) = \frac{g(b+u) - g(b) - dg(b)(u)}{\|u\|},$$

definidas en un entorno de 0 y con límite 0. Se cumple

$$h(a+v) - h(a) = dg(b)(u) + \|u\|F(u) = dg(b)\big(df(a)(v) + \|v\|E(v)\big) + \|u\|F(u)$$

$$= \Big(df(a) \circ dg\big(f(a)\big)\Big)(v) + \|v\|\,dg(b)\big(E(v)\big) + \|u\|F(u).$$

Basta probar que

$$\lim_{v \to 0} dg(b)\big(E(v)\big) + \frac{\|u\|}{\|v\|}F(u) = 0,$$

para lo cual basta a su vez probar que la función $\|u\|/\|v\|$ está acotada en un entorno de 0. Ahora bien,

$$\frac{\|u\|}{\|v\|} = \frac{\big\|df(a)(v) + \|v\|E(v)\big\|}{\|v\|} \leq \|df(a)\| + \|E(v)\|,$$

y, como E tiende a 0 en 0, está acotada en un entorno de 0. ■

Como consecuencia, $J(f \circ g)(a) = Jf(a)Jg(f(a))$.

Equivalentemente, supongamos que tenemos una función $z = z(y_1, \ldots, y_m)$, donde a su vez $y_i = y_i(x_1, \ldots, x_n)$. Entonces la regla de la cadena nos dice que, si las funciones son diferenciables,

$$\nabla z^t(x_1, \ldots, x_n) = Jy(x_1, \ldots, x_n)\nabla z^t(y_1, \ldots, y_m),$$

luego

$$\frac{\partial z}{\partial x_i} = \frac{\partial z}{\partial y_1}\frac{\partial y_1}{\partial x_i} + \cdots + \frac{\partial z}{\partial y_m}\frac{\partial y_m}{\partial x_i}.$$

Ésta es la forma explícita de la regla de la cadena.

En otros términos, si tenemos dz expresado como combinación lineal de $dy_1, \ldots, dy_m$ y cada dy_i como combinación lineal de $dx_1, \ldots, dx_n$, es decir,

$$dz = (dy_1, \ldots, dy_n)\nabla z(y_1, \ldots, y_m)^t,$$

$$(dy_1, \ldots, dy_n) = (dx_1, \ldots, dx_n)Jy(x_1, \ldots, x_n),$$

entonces, para expresar a dz como combinación lineal de $dx_1, \ldots, dx_n$ basta sustituir el segundo grupo de ecuaciones en la primera, pues así obtenemos

$$(dx_1, \ldots, dx_n)Jy(x_1, \ldots, x_n)\nabla z(y_1, \ldots, y_m)^t = (dx_1, \ldots, dx_n)\nabla z^t(x_1, \ldots, x_n),$$

es decir, $dz(x_1, \ldots, x_n)$.

Ejemplo Consideremos las funciones $z = \sqrt{x^2+y^2}$, $x = \rho\cos\theta$, $y = \rho\operatorname{sen}\theta$. Suponemos $\rho > 0$, con lo que $(x, y) \neq (0, 0)$ y todas las funciones son diferenciables (ver el teorema 4.11, más abajo). Claramente

$$\begin{aligned} dz &= \frac{x}{\sqrt{x^2+y^2}}\,dx + \frac{y}{\sqrt{x^2+y^2}}\,dy, \\ dx &= \cos\theta\,d\rho - \rho\operatorname{sen}\theta\,d\theta, \\ dy &= \operatorname{sen}\theta\,d\rho + \rho\cos\theta\,d\theta. \end{aligned}$$

Entonces

$$dz = \frac{\rho\cos\theta}{\rho}\big(\cos\theta\,d\rho - \rho\operatorname{sen}\theta\,d\theta\big) + \frac{\rho\operatorname{sen}\theta}{\rho}\big(\operatorname{sen}\theta\,d\rho + \rho\cos\theta\,d\theta\big) = d\rho,$$

que es el mismo resultado que se obtiene si diferenciamos directamente la función compuesta $z(\rho, \theta) = \rho$.

Es importante comprender que el paso del primer grupo de ecuaciones a la expresión de dz en función de ρ y θ no es una mera manipulación algebraica, sino que se fundamenta en la regla de la cadena. En este caso particular, lo que dice la regla de la cadena es:

> Si llamamos $z(\rho, \theta)$ a la función que resulta de sustituir x e y en $z(x, y)$ por sus valores en función de ρ y θ, entonces la diferencial de esta función es la que resulta de sustituir x, y, dx, dy en $dz(x, y)$ por sus valores en función de ρ, θ, $d\rho$, $d\theta$, respectivamente.

Y esto no es evidente en absoluto. ■

El teorema siguiente es el único criterio de diferenciabilidad que necesitaremos en la práctica:

Teorema 4.11 *Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$, donde A es un abierto en $\mathbb{R}^n$. Si f tiene derivadas parciales continuas en A entonces f es diferenciable en A.*

DEMOSTRACIÓN: Podemos suponer $m = 1$, pues si f tiene derivadas parciales continuas en A lo mismo vale para sus funciones coordenadas, y si éstas son diferenciables f también lo es.

Sea $a \in A$. Vamos a probar que f es diferenciable en a. Para ello basta probar que

$$\lim_{v\to 0} \frac{f(a+v) - f(a) - \sum\limits_{i=1}^n D_i f(a) v_i}{\|v\|} = 0,$$

lo que a su vez equivale a que, dado $\epsilon > 0$, exista un $\delta > 0$ de modo que si $\|v\| < \delta$ entonces

$$\left| f(a+v) - f(a) - \sum_{i=1}^n D_i f(a) v_i \right| < \epsilon \|v\|.$$

Por la continuidad de las derivadas parciales tenemos que existe un $\delta > 0$ tal que si $\|y - a\| < \delta$ entonces $y \in A$ y $|D_i f(y) - D_i f(a)| < \epsilon/n$ para $i = 1, \ldots, n$. Fijemos un v tal que $\|v\| < \delta$.

Definimos $F_i = f(a_1 + v_1, \ldots, a_i + v_i, a_{i+1}, \ldots, a_n)$. En particular, vemos que $f(a+v) = F_n$ y $f(a) = F_0$, luego

$$\begin{aligned} \left| f(a+v) - f(a) - \sum_{i=1}^n D_i f(a) v_i \right| &= \left| \sum_{i=1}^n (F_i - F_{i-1} - D_i f(a) v_i) \right| \\ &\leq \sum_{i=1}^n |F_i - F_{i-1} - D_i f(a) v_i|. \end{aligned}$$

Así pues, (teniendo en cuenta que $|v_i| \leq \|v\|$) basta probar que

$$|F_i - F_{i-1} - D_i f(a) v_i| < \frac{\epsilon}{n} |v_i|,$$

para $i = 1, \ldots, n$. Esto resulta de aplicar el teorema del valor medio a la función de una variable dada por

$$g_i(t) = f(a_1 + v_1, \ldots, a_{i-1} + v_{i-1}, a_i + t v_i, a_{i+1}, \ldots, a_n).$$

Esta función está definida en un intervalo abierto que contiene al intervalo $[0, 1]$, y el hecho de que f tenga derivadas parciales implica que g_i es derivable en su dominio. En particular es derivable en $]0, 1[$ y continua en $[0, 1]$. Además, es claro que

$$g_i'(t) = D_i f(a_1 + v_1, \ldots, a_{i-1} + v_{i-1}, a_i + t v_i, a_{i+1}, \ldots, a_n) v_i.$$

El teorema del valor medio nos da que existe $0 < t_0 < 1$ tal que

$$F_i - F_{i-1} = g_i(1) - g_i(0) = g_i'(t_0)(1 - 0).$$

Notemos que $y = (a_1 + v_1, \ldots, a_{i-1} + v_{i-1}, a_i + t_0 v_i, a_{i+1}, \ldots, a_n)$ cumple

$$\|y - a\| = \|(v_1, \ldots, v_{i-1}, t_0 v_i, 0, \ldots, 0)\| \leq \|v\| < \delta,$$

luego

$$|F_i - F_{i-1} - D_i f(a) v_i| = |D_i f(y) - D_i f(a)|\,|v_i| < \frac{\epsilon}{n}\,|v_i|,$$

como había que probar. ■

Definición 4.12 Supongamos que una función $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ admite derivada parcial respecto a una variable x_i en todo el abierto A. Si a su vez la función $D_i f$ admite derivada parcial respecto a la variable x_j en A, a esta derivada se la representa por $D_{ij} f$. También se usa la notación

$$D_{ij} f = \frac{\partial^2 f}{\partial x_i \partial x_j}.$$

Cuando el índice es el mismo se escribe

$$D_{ii} f = \frac{\partial^2 f}{\partial x_i^2}.$$

Las funciones $D_{ij} f$ se llaman *derivadas segundas* de f. Más generalmente, la función $D_{i_1 \cdots i_k} f$ será la función que resulta de derivar f respecto de i_1, derivar dicha parcial respecto a i_2, etc. Alternativamente,

$$\frac{\partial^k f}{\partial x_{i_1}^{k_1} \cdots \partial x_{i_r}^{k_r}},$$

donde $k_1 + \cdots + k_r = k$, representará la función que resulta de derivar k_1 veces f respecto a i_1, luego k_2 veces la función resultante respecto a i_2, etc. Estas funciones de llaman *derivadas parciales de orden* k de la función f.

Diremos que f es de clase C^k en A si existen todas sus derivadas parciales de orden k en A y todas ellas son continuas en A. En particular, las funciones de clase C^0 son las funciones continuas.

Obviamente una función es de clase C^{k+1} si y sólo si todas sus derivadas parciales son de clase C^k. El teorema anterior afirma que todas las funciones de clase C^1 son diferenciables. Si una función f es de clase C^2, entonces su derivadas parciales son de clase C^1, luego son diferenciables, luego son continuas y por lo tanto f es también de clase C^1. Por el mismo argumento se prueba en general que si $k \leq r$ entonces toda función de clase C^r es de clase C^k.

Las reglas de derivación justifican inmediatamente que la suma y el producto por un escalar de funciones de clase C^k es una función de clase C^k. El producto de funciones de clase C^k (con valores en $\mathbb{R}$) es de clase C^k. Lo mismo vale para el cociente si exigimos que el denominador no se anule.

Ejercicio: Probar que la composición de dos funciones de clase C^k es de clase C^k.

Ahora vamos a probar un teorema muy importante sobre derivadas sucesivas, pues nos dice que el orden de derivación no importa:

Teorema 4.13 (Teorema de Schwarz) *Si* $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ *es una función de clase* C^2 *en el abierto* A*, entonces* $D_{ij}f(a) = D_{ji}f(a)$*.*

DEMOSTRACIÓN: No perdemos generalidad si suponemos $m = 1$. También podemos suponer que $n = 2$, pues en general podemos trabajar con la función $F(x_i, x_j) = f(a_1, \ldots, x_i, \ldots, x_j, \ldots a_n)$.

Así pues, probaremos que $D_{12}f(a) = D_{21}f(a)$. Sea $a = (a_1, a_2)$. Consideremos la función

$$\Delta f(h) = f(a_1 + h, a_2 + h) - f(a_1 + h, a_2) - f(a_1, a_2 + h) + f(a_1, a_2),$$

definida en un entorno de 0. Vamos a probar que

$$D_{12}f(a) = \lim_{h\to 0^+} \frac{\Delta f(h)}{h^2}.$$

Por simetría este límite será también $D_{21}f(a)$ y el teorema estará probado.

Dado $\epsilon > 0$, la continuidad de $D_{12}f$ en a implica que existe un $h_0 > 0$ tal que si $0 < k, k' < h_0$, entonces $|D_{12}f(a_1 + k, a_2 + k') - D_{12}f(a_1, a_2)| < \epsilon$.

Podemos tomar h_0 suficientemente pequeño para que si $0 < h < h_0$ la función

$$G(t) = f(t, a_2 + h) - f(t, a_2)$$

esté definida en el intervalo $[a_1, a_1 + h]$. Por el teorema del valor medio existe un número $0 < k < h$ tal que

$$\begin{aligned} \Delta f(h) &= G(a_1 + h) - G(a_1) = hG'(a_1 + k) \\ &= h\big(D_1f(a_1 + k, a_2 + h) - D_1f(a_1 + k, a_2)\big). \end{aligned}$$

Ahora aplicamos el teorema del valor medio a la función

$$H(t) = D_1f(a_1 + k, t)$$

en el intervalo $[a_2, a_2 + h]$, que nos da un número $0 < k' < h$ tal que

$$D_1f(a_1 + k, a_2 + h) - D_1f(a_1 + k, a_2) = D_{12}f(a_1 + k, a_2 + k')h.$$

En total tenemos que

$$\Delta f(h) = D_{12}f(a_1 + k, a_2 + k')h^2,$$

luego

$$\left|\frac{\Delta f(h)}{h^2} - D_{12}f(a)\right| = |D_{12}f(a_1 + k, a_2 + k') - D_{12}f(a)| < \epsilon,$$

siempre que $0 < h < h_0$. ■

El teorema de Schwarz implica claramente que al calcular cualquier derivada parcial de orden k de una función de clase C^k es irrelevante el orden en que efectuemos las derivadas.

Ahora vamos a encaminarnos a probar el teorema de la función inversa. Esencialmente se trata de probar que las inversas de las funciones biyectivas y diferenciables son diferenciables. La situación es más complicada que en el caso de una variable, pues en el capítulo anterior vimos que toda función derivable cuya derivada no se anula es monótona, mientras que no hay ningún resultado análogo para el caso de varias variables. Por ello vamos a necesitar varios resultados previos. Entre otras cosas, nos apoyaremos en el concepto de extremo relativo y su relación con la diferenciabilidad. La situación en esto sí es análoga a la de una variable.

Definición 4.14 Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ y $a \in A$. Diremos que f tiene un *mínimo relativo* en a si existe un entorno G de a tal que para todo $p \in G$ se cumple $f(p) \geq f(a)$. Similarmente se define un *máximo relativo.* Diremos que a es un *extremo relativo* si es un máximo o un mínimo relativo.

Teorema 4.15 *Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ una función diferenciable en un punto $a \in A$. Si f tiene un extremo relativo en a, entonces $df(a) = 0$.*

DEMOSTRACIÓN: Sea $v \in \mathbb{R}^n$ y consideremos la función $\phi(h) = f(a + hv)$, para un cierto vector $v \in \mathbb{R}^n$, definida en un entorno de 0. Es claro que ϕ tiene un extremo relativo en 0. Sea $g(h) = a + hv$. Por la regla de la cadena

$$0 = \phi'(0) = d\phi(0)(1) = df\big(g(0)\big)\big(dg(0)(1)\big) = df(a)(v).$$

■

Teorema 4.16 *Sea $f : \overline{B_\delta(a)} \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ una aplicación continua, diferenciable en $B_\delta(a)$, y tal que para todo $y \in B_\delta(a)$ se cumpla $|Jf(y)| \neq 0$. Supongamos además que para todo $x \in \partial B_\delta(a)$ se cumple $f(x) \neq f(a)$. Entonces $f\big[B_\delta(a)\big]$ es un entorno de $f(a)$.*

DEMOSTRACIÓN: Sea $g : \partial B_\delta(a) \longrightarrow \mathbb{R}$ la aplicación definida mediante $g(x) = \|f(x) - f(a)\|$. Por compacidad alcanza su mínimo en un punto x. Por hipótesis $m = g(x) > 0$. Tenemos, pues, que $g(z) \geq g(x)$, para todo z tal que $\|z - a\| = \delta$. Vamos a probar que $B_{m/2}\big(f(a)\big) \subset f\big[B_\delta(a)\big]$.

Sea $y \in B_{m/2}\big(f(a)\big)$. Consideremos la función $h : \overline{B_\delta(a)} \longrightarrow \mathbb{R}$ dada por $h(x) = \|f(x) - y\|$. Por compacidad alcanza su mínimo en un punto z y, puesto que $h(a) = \|f(a) - y\| < m/2$, vemos que $h(z) < m/2$.

Si $x \in \partial B_\delta(a)$, entonces

$$h(x) = \|f(x) - y\| \geq \|f(x) - f(a)\| - \|f(a) - y\| > g(x) - \frac{m}{2} \geq m - \frac{m}{2} = \frac{m}{2},$$

luego $h(x)$ no es el mínimo de h. En otros términos, $z \notin \partial B_\delta(a)$, luego $z \in B_\delta(a)$

Es claro que h^2 también alcanza su mínimo en z y

$$h^2(x) = \sum_{k=1}^{n} (f_k(x) - y_k)^2.$$

Por consiguiente,

$$D_j h^2(z) = \sum_{k=1}^{n} 2(f_k(z) - y_k) D_j f_k(z) = 0,$$

pues z es un extremo. Matricialmente tenemos $\big(f(z) - y\big)Jf(z)^t = 0$ y como el determinante de la matriz es no nulo por hipótesis, $y = f(z) \in f\big[B_\delta(a)\big]$. ■

Teorema 4.17 *Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ una aplicación inyectiva, diferenciable en el abierto A y tal que $|Jf(x)| \neq 0$ para todo $x \in A$. Entonces f es abierta, luego $f : A \longrightarrow f[A]$ es un homeomorfismo.*

DEMOSTRACIÓN: Sea U un abierto en A. Veamos que $f[U]$ es abierto en $\mathbb{R}^n$. Tomemos un punto $a \in U$ y veamos que $f[U]$ es entorno de $f(a)$. Existe un $\delta > 0$ tal que $\overline{B_\delta(a)} \subset U$, y la restricción de f a esta bola cerrada está en las hipótesis del teorema anterior. Por consiguiente $f\big[B_\delta(a)\big] \subset f[U]$ es un entorno de $f(a)$. ■

Ahora ya podemos probar el teorema de la función inversa.

Teorema 4.18 (Teorema de la función inversa) *Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ una función inyectiva de clase C^k, con $k \geq 1$, en el abierto A y tal que se cumpla $|Jf(x)| \neq 0$ para todo $x \in A$. Entonces $B = f[A]$ es abierto y $f^{-1} : B \longrightarrow A$ es de clase C^k en B.*

DEMOSTRACIÓN: Llamemos $g = f^{-1}$. Por el teorema anterior f y g son homeomorfismos. Veamos que tiene g parciales continuas. Sea e_i el i-ésimo vector de la base canónica de $\mathbb{R}^n$. Sea $b \in B$ y $a = g(b)$. Tomemos una bola abierta $B_\eta(a) \subset A$ y un $\alpha > 0$ suficientemente pequeño tal que el segmento de extremos b y $b + \alpha e_i$ esté contenido en $f[B_\eta(a)]$. Sea $a' = g(b + \alpha e_i)$. Entonces $a' \in B_\eta(a)$, luego el segmento de extremos a y a' está contenido en A.

Claramente $f(a') - f(a) = \alpha e_i$. Si f_j es la j-ésima función coordenada de f, tenemos

$$f_j(a') - f_j(a) = \alpha\delta_{ij},$$

donde (δ_{ij}) es la matriz identidad. Aplicamos el teorema del valor medio a la función $\phi(t) = f_j\big(a + t(a' - a)\big)$, definida en $[0, 1]$, en virtud del cual existe $0 < t < 1$ tal que

$$\alpha\delta_{ij} = \phi(1) - \phi(0) = d\phi(t)(1) = df_j\big(a + t(a' - a)\big)(a' - a).$$

Sea $z^j = a + t(a' - a)$. Así

$$\alpha\delta_{ij} = \sum_{k=1}^{n} D_k f_j(z^j)(a'_k - a_k).$$

Matricialmente tenemos

$$\alpha I = (a' - a)\big(D_k f_j(z^j)\big).$$

La función $h : A^n \longrightarrow \mathbb{R}$ dada por $h(z^1, \ldots, z^n) = \bigl|(D_k f_j(z^j))\bigr|$ es continua, pues las derivadas parciales son continuas y el determinante es un polinomio. Por hipótesis tenemos que $h(a, \ldots, a) \neq 0$, luego existe un entorno de $(a, \ldots, a)$ en el cual h no se anula. Tomando α suficientemente pequeño podemos exigir que $(a', \ldots, a')$ esté en una bola de centro $(a, \ldots, a)$ contenida en dicho entorno, con lo que el punto $(z^1, \ldots, z^n)$ que hemos construido también está en dicho entorno, luego $\bigl|(D_k f_j(z^j))\bigr| \neq 0$ y podemos calcular la matriz inversa, cuyos coeficientes se expresan como un cociente de determinantes de matrices cuyos coeficientes son derivadas parciales. En definitiva obtenemos una expresión de la forma

$$\frac{a'_k - a_k}{\alpha} = \frac{P_k\bigl(D_k f_j(z^j)\bigr)}{Q_k\bigl(D_k f_j(z^j)\bigr)},$$

donde P_k y Q_k son polinomios. Por definición de a y a' tenemos

$$\frac{g_k(b + \alpha e_i) - g_k(b)}{\alpha} = \frac{P_k\bigl(D_k f_j(z^j)\bigr)}{Q_k\bigl(D_k f_j(z^j)\bigr)}.$$

Tomando α suficientemente pequeño podemos conseguir que $\|z^i - a\|$ se haga arbitrariamente pequeño. Por la continuidad de las derivadas parciales podemos exigir que $|D_k f_j(z^j) - D_k f_j(a)|$ se haga arbitrariamente pequeño y por la continuidad de los polinomios P_k y Q_k podemos hacer que el miembro derecho de la ecuación anterior se aproxime cuanto queramos al término correspondiente con a en lugar de los puntos z^i. En definitiva, existe

$$D_i g_k(b) = \lim_{\alpha \to 0} \frac{g_k(b + \alpha e_i) - g_k(b)}{\alpha} = \frac{P_k\bigl(D_k f_j(g(b))\bigr)}{Q_k\bigl(D_k f_j(g(b))\bigr)}.$$

Más aún, los polinomios P_k y Q_k son los que expresan las soluciones de una ecuación lineal en términos de sus coeficientes, luego no dependen de b, luego esta expresión muestra también que $D_i g_k$ es una composición de funciones continuas, luego es continua. Más en general, si f es de clase C^k, entonces g también lo es. ■

Si f está en las condiciones del teorema anterior tenemos que $f \circ f^{-1} = f^{-1} \circ f = 1$, luego por la regla de la cadena

$$df(a) \circ df^{-1}(b) = df^{-1}(b) \circ df(a) = 1,$$

luego las dos diferenciales son biyectivas y $df^{-1}(b) = df(a)^{-1}$.

Equivalentemente, si $y_i(x_1, \ldots, x_n)$ es una transformación biyectiva y diferenciable con determinante jacobiano no nulo y llamamos $x_i(y_1, \ldots, y_n)$ a la función inversa, entonces el sistema de ecuaciones

$$\begin{aligned} dy_1 &= \frac{\partial y_1}{\partial x_1} dx_1 + \cdots + \frac{\partial y_1}{\partial x_n} dx_n \\ \vdots & \qquad \vdots \qquad\qquad \vdots \\ dy_n &= \frac{\partial y_n}{\partial x_1} dx_1 + \cdots + \frac{\partial y_n}{\partial x_n} dx_n \end{aligned}$$

tiene por matriz de coeficientes a la matriz jacobiana de la transformación, luego podemos despejar $dx_1, \ldots, dx_n$ en función de $dy_1, \ldots, dy_n$ y así obtenemos precisamente la diferencial de la función inversa.

Ejemplo Como ya advertíamos, al contrario de lo que sucede en una variable, el hecho de que una aplicación tenga diferencial no nula en todo punto (o incluso determinante jacobiano no nulo) no garantiza que sea biyectiva. Por ejemplo, consideremos $f(x,y) = (e^x \cos y, e^x \operatorname{sen} y)$ (vista como aplicación de $\mathbb{C}$ en $\mathbb{C}$, se trata simplemente de la exponencial compleja). La matriz jacobiana de f es

$$\begin{pmatrix} e^x \cos y & e^x \operatorname{sen} y \\ -e^x \operatorname{sen} y & e^x \cos y \end{pmatrix}$$

y su determinante en cada punto (x,y) es $e^x \neq 0$. Por otro lado es fácil ver que f no es biyectiva. ■

Lo máximo que podemos deducir del hecho de que el determinante jacobiano no se anule es que la función es localmente inyectiva. La prueba se basa en un argumento que hemos usado en la prueba del teorema de la función inversa.

Teorema 4.19 (Teorema de inyectividad local) *Sea $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ una función de clase C^1 en el abierto A y sea $a \in A$ tal que $|Jf(a)| \neq 0$. Entonces existe un entorno B de a tal que $|Jf(x)| \neq 0$ para todo $x \in B$ y f es inyectiva sobre B.*

Demostración: La función $h : A^n \longrightarrow \mathbb{R}$ dada por

$$h(z^1, \ldots, z^n) = \big|(D_k f_j(z^j))\big|$$

es continua, pues las derivadas parciales son continuas y el determinante es un polinomio. Por hipótesis tenemos que $h(a, \ldots, a) \neq 0$, luego existe un entorno de $(a, \ldots, a)$ en el cual h no se anula. Este entorno lo podemos tomar de la forma $B_\delta(a) \times \cdots \times B_\delta(a)$. Tomaremos $B = B_\delta(a)$.

Veamos que si x, $y \in B_\delta(a)$ y $f(x) = f(y)$ entonces $x = y$. Consideremos la función $\phi(t) = f_i(x + t(y - x))$, definida en $[0,1]$. Claramente es derivable. Por el teorema del valor medio,

$$f_i(y) - f_i(x) = d\phi(t_i)(1) = df_i(x + t_i(y - x))(y - x) = df_i(z^i)(y - x),$$

donde z^i es un punto entre x e y, luego $z^i \in B_\delta(a)$. Por lo tanto

$$f_i(y) - f_i(x) = \sum_{k=1}^{n} D_k f_i(z^i)(y_k - x_k),$$

lo que matricialmente se expresa como

$$0 = f(y) - f(x) = (y - x)(D_k f_i(z^i)).$$

La matriz tiene determinante $h(z^1, \ldots, z^n) \neq 0$, luego ha de ser $x = y$. ■

Para terminar generalizamos a varias variables un hecho que tenemos probado para el caso de una:

Teorema 4.20 *Si $f : A \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ es una función diferenciable en un abierto conexo A y $df = 0$, entonces f es constante.*

DEMOSTRACIÓN: Podemos suponer $m = 1$. Dados dos puntos a, $b \in A$, existe una poligonal contenida en A con extremos a y b. Basta probar que f toma el mismo valor en los vértices de la poligonal, luego en definitiva basta probar que si a y b son los extremos de un segmento contenido en A entonces $f(a) = f(b)$. Consideramos la función $\phi(t) = f(a + t(b - a))$ en $[0, 1]$ y le aplicamos el teorema del valor medio. Concluimos que

$$f(b) - f(a) = \phi'(t) = df(a + t(b - a))(b - a) = 0.$$

■

4.3 Curvas parametrizables

Las técnicas de este capítulo nos capacitan para estudiar curvas más eficientemente que las del capítulo anterior. En efecto, allí estudiábamos curvas considerándolas como gráficas de funciones de una variable, pero esto no permite trabajar con curvas cualesquiera. Por ejemplo, una elipse no es la gráfica de ninguna función. Ahora podemos aplicar la derivabilidad al estudio de curvas en el sentido de aplicaciones $x : I \longrightarrow \mathbb{R}^n$, donde I es un intervalo en $\mathbb{R}$. Para que una tal aplicación x pueda ser llamada "curva" razonablemente, deberemos exigir al menos que sea continua. Aquí nos centraremos en las curvas que además son derivables. Si una curva está definida en un intervalo cerrado $[a, b]$, exigiremos que sea derivable en $]a, b[$. A estas curvas las llamaremos *arcos*. Si $x : [a, b] \longrightarrow \mathbb{R}^n$, los puntos $x(a)$ y $x(b)$ se llaman *extremos* del arco. Concretamente, $x(a)$ es el *extremo inicial* y $x(b)$ es el *extremo final*. La gráfica de una función continua $f : [a, b] \longrightarrow \mathbb{R}$ puede identificarse con el arco $x(t) = \big(t, x(t)\big)$. De este modo, el tratamiento de los arcos que estamos dando aquí generaliza al del capítulo anterior.

Según lo dicho, una curva no es un mero conjunto de puntos en $\mathbb{R}^n$, sino un conjunto de puntos recorridos de un modo en concreto. Si $x(t)$ es una curva, la variable t se llama *parámetro* de la misma. Conviene imaginarse a t como una variable temporal, de modo que $x(t)$ es la posición en el instante t de un punto móvil que recorre la curva. La imagen de x es la trayectoria del móvil.

Vamos a interpretar la derivabilidad de una curva $x(t)$ en un punto t. Si existe $x'(t) = v \neq 0$, entonces para valores pequeños de h tenemos

$$\frac{x(t+h) - x(t)}{h} \approx v, \tag{4.1}$$

luego $x(t+h) \approx x(t)+hv$. La curva $x(t)+hv$, cuando varía h, recorre los puntos de una recta, y estamos diciendo que para valores pequeños de h la curva x se parece a dicha recta. Así pues, la recta de dirección $x'(t)$ se confunde con la curva alrededor de $x(t)$, y por ello la llamamos *recta tangente* a x en $x(t)$. Esto

interpreta la dirección de $x'(t)$. Es fácil ver que su sentido es el sentido de avance al recorrer la curva. Observemos que si $f : I \longrightarrow \mathbb{R}$ es una función derivable en un punto t, entonces la tangente del arco $x(t) = \big(t, f(t)\big)$ es la recta de dirección $\big(1, f'(t)\big)$, luego su pendiente es $f'(t)$, luego coincide con la tangente tal y como la definimos en el capítulo anterior.

Ya tenemos interpretados la dirección y el sentido de $x'(t)$. Falta interpretar su módulo. Claramente se trata del límite del módulo de (4.1). La cantidad $\|x(t+h) - x(t)\|$ es la distancia que recorremos en h unidades de tiempo desde el instante t, luego al dividir entre $|h|$ obtenemos la distancia media recorrida por unidad de tiempo en el intervalo de extremos t y $t+h$, es decir, la velocidad media en dicho intervalo. El límite cuando $h \to 0$ es, pues, la velocidad con que recorremos la curva en el instante t. En realidad los físicos prefieren llamar *velocidad* a todo el vector $x'(t)$, de modo que la dirección indica hacia dónde nos movemos en el instante t y el módulo indica con qué rapidez lo hacemos.

Conviene precisar estas ideas. La *primera ley de Newton* afirma que si un cuerpo está libre de toda acción externa, permanecerá en reposo o se moverá en línea recta a velocidad constante. Todo esto se resume en que su velocidad en sentido vectorial permanece constante.

Ejemplo Consideremos la curva $x(t) = (\cos t, \operatorname{sen} t)$, definida en el intervalo $[0, +\infty[$. Esta curva recorre infinitas veces la circunferencia de centro $(0, 0)$ y radio 1. Su velocidad es $x'(t) = (-\operatorname{sen} t, \cos t)$, cuyo módulo es constante igual a 1, esto significa que recorremos la circunferencia a la misma velocidad, digamos de un metro por segundo. Para que un objeto siga esta trayectoria es necesario que una fuerza lo obligue a mantenerse a la misma distancia del centro. Por ejemplo, el móvil podría ser un cuerpo que gira atado a una cuerda. El hecho de que $x'(2\pi) = (0, 1)$ significa que si en el instante 2π se cortara la cuerda entonces el cuerpo, libre ya de la fuerza que le retenía, saldría despedido hacia arriba a razón de un metro por segundo.

Consideremos ahora la curva $x(t) = (\cos t^2, \operatorname{sen} t^2)$. Su trayectoria es la misma, pero ahora la velocidad es $x'(t) = (-2t \operatorname{sen} t^2, 2t \cos t^2)$, cuyo módulo es $2t$, lo que significa que ahora el móvil gira cada vez más rápido. Comienza a velocidad 0, al dar una vuelta alcanza la velocidad de 2 metros por segundo, a la segunda vuelta de 4, etc. ■

Las consideraciones anteriores muestran que la derivabilidad de una curva se traduce en la existencia de una recta tangente salvo que la derivada sea nula. La existencia de una recta tangente en $x(t)$ significa que el arco se confunde con una recta alrededor de $x(t)$, con lo que el arco no puede formar un "pico" en $x(t)$. Esto ya no es cierto si $x'(t) = 0$. Por ejemplo, la curva $x(t) = (t^3, |t^3|)$ es derivable en 0, pues su derivada por la izquierda coincide con la de $(t^3, -t^3)$ y su derivada por la derecha coincide con la de (t^3, t^3), y ambas son nulas, pero su gráfica es la misma que la de $(t, |t|)$, es decir, la de la función $|x|$, que tiene un pico en 0. Esto nos lleva a descartar las curvas con derivada nula.

Definición 4.21 Una *curva parametrizada regular* $x : I \longrightarrow \mathbb{R}^n$ es una aplicación definida sobre un intervalo abierto $I \subset \mathbb{R}$ derivable y con derivada no nula en ningún punto.

El vector
$$T(t) = \frac{x'(t)}{\|x'(t)\|}$$
se llama *vector tangente* a x en el punto $x(t)$. La recta que pasa por $x(t)$ con dirección $T(t)$ se llama *recta tangente* a x por $x(t)$.

Hemos visto un ejemplo de una misma trayectoria recorrida a velocidades distintas. Desde un punto de vista geométrico, lo que importa de una curva es su forma, y no la velocidad con la que se recorre. Vamos a explicitar esta distinción.

Dado un arco parametrizado $x : [a, b] \longrightarrow \mathbb{R}^n$, un *cambio de parámetro* es una aplicación $t : [u, v] \longrightarrow [a, b]$ que se extiende a un intervalo abierto mayor donde es derivable y la derivada no se anula. Por consiguiente t es biyectiva y t' tiene signo constante. Diremos que t es un cambio de parámetro *directo* o *inverso* según si $t' > 0$ o $t' < 0$. El arco parametrizado $y(s) = x(t(s))$ se llama *reparametrización* de x mediante el cambio de parámetro t.

Diremos que dos arcos parametrizados regulares x e y son *(estrictamente) equivalentes* si existe un cambio de parámetro (directo) que transforma uno en otro.

Es claro que la identidad es un cambio de parámetro directo, la inversa de un cambio de parámetro (directo) es un cambio de parámetro (directo) y la composición de dos cambios de parámetro (directos) es un cambio de parámetro (directo). De aquí se sigue que la equivalencia y la equivalencia estricta son relaciones de equivalencia entre los arcos parametrizados.

Llamaremos *arcos regulares* a las clases de equivalencia estricta de arcos parametrizados regulares, de modo que dos elementos de la misma clase se considerarán dos *parametrizaciones* de un mismo arco.

Todas las parametrizaciones de un arco tienen la misma imagen, a la que podemos llamar *imagen* del arco. Los cambios de parámetro directos son crecientes, por lo que dos parametrizaciones de un mismo arco tienen los mismos extremos, a los que podemos llamar *extremos* del arco.

Consideremos dos parametrizaciones $x : [a, b] \longrightarrow \mathbb{R}^n$, $y : [c, d] \longrightarrow \mathbb{R}^n$ de un mismo arco. Entonces $y(s) = x(t(s))$ para un cierto cambio de parámetro directo t. Consideremos ahora dos cambios de parámetro inversos
$$u : [a', b'] \longrightarrow [a, b], \quad v : [c', d'] \longrightarrow [c, d]$$
y las reparametrizaciones $x(u(r))$, $y(v(r))$. Entonces $y(v(r)) = x(t(v(r)) = x(u(u^{-1}(t(v(r)))))$ y $v \circ t \circ u^{-1}$ es un cambio de parámetro directo, luego las dos reparametrizaciones corresponden a un mismo arco. Por lo tanto, podemos definir el *arco inverso* de un arco x al arco resultante de componer con un cambio de parámetro inverso cualquiera de las parametrizaciones de x. Lo

representaremos por $-x$. Es claro que x y $-x$ tienen la misma imagen, pero el extremo inicial de x es el extremo final de $-x$ y viceversa.

De este modo, la noción de arco como clase estricta de arcos parametrizados recoge el concepto geométrico de arco regular independiente del modo en que se recorre, pero conservando el sentido del recorrido. Si consideramos clases no estrictas identificamos cada arco con su inverso, y con ello hacemos abstracción incluso del sentido en que lo recorremos. Todos estos conceptos se aplican igualmente a curvas cualesquiera.

Longitud de un arco Consideremos el arco $x(t) = (r\cos t, r\,\mathrm{sen}\,t)$, con $r > 0$. Su derivada tiene módulo r, lo que se interpreta como que el arco recorre la circunferencia de centro $(0,0)$ y radio r a una velocidad constante de r unidades de longitud por unidad de tiempo. Puesto que recorre la circunferencia completa en 2π unidades de tiempo, concluimos que el espacio que recorre, es decir, la longitud de la circunferencia, es $2\pi r$. Más en general, mediante esta parametrización recorremos un arco de α radianes en α unidades de tiempo, luego la longitud de un arco de α radianes es αr, tal y como anticipábamos en el capítulo anterior. Vamos a generalizar este argumento para definir la longitud de un arco arbitrario.

Sea $x : [a,b] \longrightarrow \mathbb{R}^n$ un arco y vamos a definir la función $s : [a,b] \longrightarrow \mathbb{R}$ tal que $s(t)$ es la longitud de arco entre $x(a)$ y $x(t)$. Obviamente ha de cumplir $s(a) = 0$. Supongamos que es derivable y vamos a calcular su derivada en un punto t. El arco x se confunde con una recta alrededor de $x(t)$. Esto significa que si h es suficientemente pequeño el arco entre $x(t)$ y $x(t+h)$ es indistinguible del segmento que une ambos puntos, luego tendremos

$$s(t+h) - s(t) \approx \pm\|x(t+h) - x(t)\|,$$

donde el signo es el de h. La aproximación será mejor cuanto menor sea h. Dividiendo entre h queda

$$\frac{s(t+h) - s(t)}{h} \approx \left\|\frac{x(t+h) - x(t)}{h}\right\|.$$

Estos dos cocientes se parecerán más cuanto menor sea h, lo que se traduce en que los límites cuando $h \to 0$ han de coincidir. Así:

$$\frac{ds}{dt} = \|x'(t)\|, \tag{4.2}$$

luego

$$s(t) = \int_a^t \|x'(u)\|\,du.$$

En particular, la longitud del arco completo será

$$L(x) = \int_a^b \|x'(t)\|\,dt.$$

Definición 4.22 Diremos que un arco parametrizado regular $x : [a, b] \longrightarrow \mathbb{R}^n$ es *rectificable* si la función $\|x'(t)\|$ tiene primitiva en $[a, b]$ (es decir, si tiene primitiva en el intervalo abierto y ésta se extiende continuamente al intervalo cerrado), y entonces llamaremos *longitud* de x al número real

$$L(x) = \int_a^b \|x'(t)\| \, dt.$$

Notar que como el integrando es positivo, su primitiva ha de ser estrictamente creciente, luego $L(x) > 0$. La longitud de un arco parametrizado coincide con la de cualquiera de sus reparametrizaciones. En efecto, si $y(s) = x(t(s))$, entonces $y'(s) = x'(t(s))t'(s)$, luego

$$L(x) = \int_a^b \|x'(t)\| \, dt = \int_{s(a)}^{s(b)} \|x'(t(s))\| t'(s) \, ds = \int_{s(a)}^{s(b)} \|y'(s)\| \, ds = L(y).$$

Ejercicio: Comprobar que la longitud de un arco coincide con la de su opuesto, y que la longitud es invariante por isometrías.

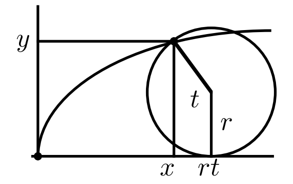
Ahora es inmediato comprobar que la longitud de un arco de circunferencia de radio r y amplitud α es α. En particular la longitud de una circunferencia de radio r es $2\pi r$. De hecho, los griegos llamaron π a esta constante por ser la proporción entre el *perímetro* de la circunferencia y su diámetro.

Sea $x : [a, b] \longrightarrow \mathbb{R}^n$ un arco rectificable y sea $s(t)$ la longitud de arco entre $x(a)$ y $x(t)$. Hemos visto que se trata de una función derivable y que $s'(t) = \|x'(t)\|$. Por lo tanto s es creciente y biyectiva. Su inversa $t : [0, L] \longrightarrow [a, b]$ es un cambio de parámetro, la función $x(s) = x(t(s))$ es una reparametrización del arco y

$$x'(s) = x'(t)t'(s) = \frac{x'(t)}{\|x'(t)\|},$$

luego $\|x'(s)\| = 1$ y así $x'(s) = T(s)$. Más concretamente, $x(s)$ se caracteriza por que la longitud de arco entre $x(0)$ y $x(s)$ es s. A esta parametrización del arco la llamaremos *parametrización natural.* También diremos entonces que x está *parametrizado por el arco* Desde un punto de vista cinemático, la parametrización natural se interpreta como la que recorre el arco a velocidad constante igual a 1 (constante en módulo).

Ejemplo La trayectoria de un clavo de una rueda que gira se conoce con el nombre de *cicloide.* Vamos a calcular la longitud de una vuelta de cicloide.

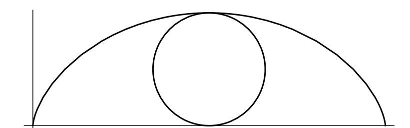


Primeramente necesitamos encontrar una parametrización de la curva. Supongamos que la rueda gira sobre el eje $y = 0$ y que el clavo parte de la posición $(0, 0)$. Si llamamos r al radio de la circunferencia, vemos que cuando la rueda ha girado t radianes su centro se encuentra en el punto (rt, r), luego el clavo se encuentra en $x(t) = (rt - r \operatorname{sen} t, r - r \cos t)$.

Por consiguiente $x'(t) = r(1 - \cos t, \operatorname{sen} t)$,

$$\|x'(t)\| = r\sqrt{2 - 2\cos t} = 2r \operatorname{sen} \frac{t}{2}.$$

Vemos que los múltiplos de 2π son puntos singulares de la cicloide. Corresponden a los momentos en que el clavo toca el suelo. Entonces se para y cambia de sentido.

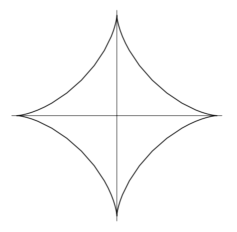


Es fácil calcular

$$L = \int_0^{2\pi} 2r \operatorname{sen} \frac{t}{2}\, dt = 4r \left[-\cos \frac{t}{2}\right]_0^{2\pi} = 8r.$$

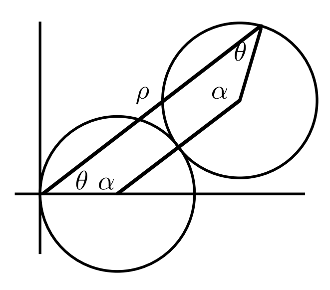
■

Ejercicio: Calcular la longitud de la *astroide,* dada por $x(t) = (a\cos^3 t, a\operatorname{sen}^3 t)$.



Ejemplo La trayectoria de un clavo de una rueda que gira sobre una circunferencia se llama *epicicloide.* El caso más simple lo tenemos cuando ambas circunferencias tienen el mismo radio. La curva se llama entonces *cardioide,* porque su forma recuerda a un corazón.

Supongamos que la circunferencia fija tiene centro en $(a/4, 0)$ y radio $a/4$ y que el clavo parte de la posición $(0, 0)$. Cuando la rueda ha girado α radianes la situación es la que indica la figura. El cuadrilátero tiene dos ángulos y dos lados iguales, por lo que los otros dos ángulos también tienen la misma amplitud θ. Es claro entonces que

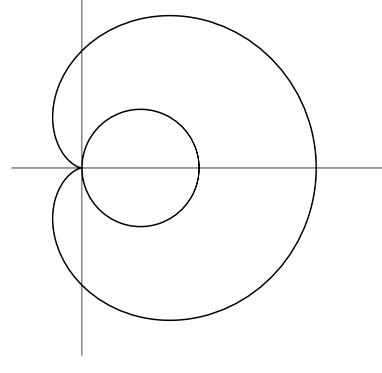


$$\rho = \frac{a}{2} + 2\frac{a}{4}\cos\theta,$$

luego

$$\rho = \frac{a}{2}(1 + \cos\theta),$$

donde $\theta \in\,]-\pi, \pi[$. Ésta es la ecuación de la cardioide en coordenadas polares.



Vamos a ver en general cuál es la expresión de la longitud de una curva parametrizada en coordenadas polares $\bigl(\rho(t), \theta(t)\bigr)$. Notemos que (4.2), para el caso de dos variables puede escribirse también como

$$ds^2 = dx^2 + dy^2.$$

Se dice que ésta es la expresión del elemento de longitud (es decir, de una longitud infinitesimal) en coordenadas cartesianas. Se trata de la versión infinitesimal del teorema de pitágoras. Si diferenciamos las relaciones $x = \rho \cos\theta$, $y = \rho \operatorname{sen}\theta$ obtenemos

$$dx = \cos\theta\, d\rho - \rho \operatorname{sen}\theta\, d\theta, \quad dy = \operatorname{sen}\theta\, d\rho + \rho \cos\theta\, d\theta.$$

Sustituyendo queda

$$ds^2 = d\rho^2 + \rho^2\, d\theta^2. \tag{4.3}$$

Ésta es la expresión del elemento de longitud de un arco en coordenadas polares. Aplicado a la cardioide resulta

$$ds^2 = \frac{a^2}{2}(1 + \cos\theta)\, d\theta^2,$$

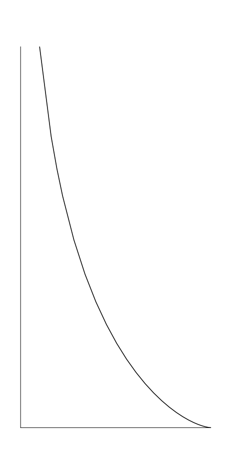
de donde

$$ds = a \cos\frac{\theta}{2}\, d\theta.$$

Así pues, la longitud de la cardioide es

$$L = \int_{-\pi}^{\pi} a \cos\frac{\theta}{2}\, d\theta = 2a \left[\operatorname{sen}\frac{\theta}{2}\right]_{-\pi}^{\pi} = 4a.$$

■



Ejemplo Consideremos un cuerpo puntual situado en $(l, 0)$ atado a una cuerda de longitud l con su otro extremo en $(0, 0)$. Estiramos de la cuerda de modo que su extremo suba por el eje Y. La trayectoria del cuerpo arrastrado por la cuerda recibe el nombre de *tractriz*. Vamos a obtener una parametrización de la tractriz. Un cuerpo estirado por una cuerda se mueve en la dirección de la cuerda, luego ésta ha de ser tangente a la trayectoria. Si llamamos $y = f(x)$ a la tractriz, definida para $0 < x < l$, entonces su recta tangente es

$$Y = f(x) + f'(x)(X - x).$$

Cortará al eje Y en el punto $\big(0, f(x) - f'(x)x\big)$. Éste es el punto donde está el extremo de la cuerda cuando el otro extremo está en $\big(x, f(x)\big)$. La distancia entre ambos debe ser, pues, igual a l. Por consiguiente:

$$x^2 + f'(x)^2 x^2 = l^2.$$

Despejando obtenemos

$$dy = -\sqrt{\frac{l^2}{x^2} - 1}\, dx.$$

El signo negativo se debe a que, tal y como hemos planteado el problema, la función y ha de ser decreciente. El cambio $x = l \operatorname{sen} \theta$ biyecta los números $0 < x \leq l$ con los números $\pi/2 \leq \theta < \pi$. La igualdad anterior se transforma en

$$dy = -l\sqrt{\frac{1}{\operatorname{sen}^2 \theta} - 1}\cos\theta\, d\theta = l\,\frac{\cos^2\theta}{\operatorname{sen}\theta}\, d\theta = \frac{l\, d\theta}{\operatorname{sen}\theta} - l \operatorname{sen}\theta\, d\theta.$$

La posición inicial corresponde a $x = l$, luego a $\theta = \pi/2$, es decir, $y(\pi/2) = 0$, luego

$$y(\theta) = l\int_{\pi/2}^{\theta} \frac{dt}{\operatorname{sen} t} - l\int_{\pi/2}^{\theta} \operatorname{sen} t\, dt.$$

Existen reglas de integración que permiten calcular metódicamente la primera primitiva. Como no nos hemos ocupado de ellas nos limitaremos a dar el resultado. El lector puede comprobar sin dificultad que es correcto derivando la solución que presentamos.

$$y(\theta) = l\left[\log\tan\frac{t}{2}\right]_{\pi/2}^{\theta} + l\,[\cos t]_{\pi/2}^{\theta} = l\log\tan\frac{\theta}{2} + l\cos\theta.$$

Así pues, la tractriz viene dada por

$$T(\theta) = (l\operatorname{sen}\theta, l\log\tan\frac{\theta}{2} + l\cos\theta), \quad \theta \in [\pi/2, \pi[\,.$$

Su derivada es

$$T'(\theta) = \left(l\cos\theta, l\,\frac{\cos^2\theta}{\operatorname{sen}\theta}\right), \quad \|T'(\theta)\| = -l\,\frac{\cos\theta}{\operatorname{sen}\theta}.$$

La tractriz es regular en $]\pi/2, \pi[$. La longitud de un arco de tractriz es

$$s(\theta) = -l\int_{\pi/2}^{\theta} \frac{\cos t}{\operatorname{sen} t}\,dt = -l\,[\log\operatorname{sen} t]_{\pi/2}^{\theta} = -l\log\operatorname{sen}\theta.$$

■

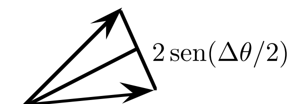
Vector normal y curvatura Sea $x(s)$ un arco parametrizado por la longitud de arco. Supongamos que admite derivada segunda. Derivando la igualdad $x'(s)x'(s) = 1$ obtenemos que $2x''(s)x'(s) = 0$, luego $x''(s) \perp x'(s)$. Supuesto que $x''(s) \neq 0$ podemos definir el *vector normal* del arco como

$$N(s) = \frac{x''(s)}{\|x''(s)\|},$$

y la *curvatura* del arco como $\kappa(s) = \|x''(s)\|$.

Para interpretar la curvatura llamemos $\Delta\theta$ al ángulo entre los vectores $x'(s)$ y $x'(s+\Delta s)$, donde $\Delta\theta$ es una función de Δs. Puesto que x' tiene módulo constante igual a 1, la trigonometría nos da que

$$\|x'(s+\Delta s) - x'(s)\| = 2\operatorname{sen}\frac{\Delta\theta}{2}.$$



Por consiguiente,

$$\frac{\|x'(s+\Delta s) - x'(s)\|}{|\Delta s|} = \frac{\operatorname{sen}\frac{\Delta\theta}{2}}{\Delta\theta/2}\,\frac{\Delta\theta}{|\Delta s|}.$$

Es claro que $\Delta\theta \to 0$ cuando $\Delta s \to 0$, luego

$$\kappa(s) = \lim_{\Delta s\to 0} \frac{\Delta\theta}{|\Delta s|}.$$

Así pues, la curvatura mide la variación del ángulo del vector tangente por unidad de arco recorrido. Es claro que cuanto mayor sea la curvatura "más curvado" estará el arco.

Ejemplo La parametrización natural de una circunferencia de radio r es $x(s) = \big(r\cos(s/r), r\operatorname{sen}(s/r)\big)$. El vector tangente es, por lo tanto, $x'(s) = \big(-\operatorname{sen}(s/r), \cos(s/r)\big)$. De aquí obtenemos

$$x''(s) = \left(-\frac{1}{r}\,\cos\frac{s}{r}, -\frac{1}{r}\,\operatorname{sen}\frac{s}{r}\right),$$

con lo que el vector normal es $N(s) = -\big(\cos(s/r), \operatorname{sen}(s/r)\big)$ y la curvatura es $\kappa(s) = 1/r$. Intuitivamente es claro que una circunferencia está menos curvada cuanto mayor es su radio, tal y como se pone de manifiesto en la fórmula que hemos obtenido. ■

Dada una curva x de curvatura no nula en un punto dado $x(s)$, la circunferencia de radio $r = 1/\kappa(s)$ y centro $x(s) - \kappa N(s)$ pasa por $x(s)$ y tiene el mismo vector tangente, el mismo vector normal y la misma curvatura en este punto. Esto hace que sea la circunferencia que más se parece a x en un entorno de $x(s)$ y se la llama *circunferencia osculatriz* a x por $x(s)$. El radio $r(s) = 1/\kappa(s)$ se llama *radio de curvatura* de x en $x(s)$.

Veamos ahora fórmulas explícitas para calcular el vector normal y la curvatura cuando la parametrización no es la natural. Conviene usar el lenguaje de la cinemática. Supongamos que $x(t)$ representa la posición de un móvil puntual en función del tiempo. Llamaremos $x(s)$ a la parametrización natural. Entonces la velocidad del móvil es $V = x'(t)$. Si llamamos $v = \|V\|$, entonces sabemos que $v = s'(t)$, con lo que v es la *velocidad sobre la trayectoria,* que mide la distancia recorrida por unidad de tiempo. Además $V = x'(t) = x'(s)s'(t)$, es decir,

$$V = vT.$$

Se define el vector *aceleración* como $A = V' = X''(t)$. Derivando en la relación anterior tenemos

$$A = v'(t)T(t) + v(t)T'(t).$$

Llamaremos $a(t) = v'(t)$, que es la *aceleración sobre la trayectoria,* es decir la variación de la velocidad sobre la trayectoria por unidad de tiempo. Aplicando la regla de la cadena a $T(t) = T(s(t))$ obtenemos $T'(t) = T'(s)s'(t) = \kappa vN$, luego

$$A = aT + \kappa v^2 N = aT + \frac{v^2}{r}N.$$

Vemos, pues, que la aceleración se descompone de forma natural en una componente *tangencial,* que mide la variación del módulo de la velocidad, y una componente *normal,* que determina la curvatura.

Ahora multiplicamos esta igualdad por sí misma: $\|A\|^2 = a^2 + \kappa^2 v^4$, de donde

$$\kappa^2 = \frac{\|A\|^2 - a^2}{v^4}.$$

Ahora bien,

$$a = v' = \|V\|' = \frac{VV'}{\|V\|} = \frac{VV'}{v},$$

luego

$$\kappa^2 = \frac{\|A\|^2 v^2 - (VV')^2}{v^6} = \frac{\|V \wedge A\|^2}{v^6},$$

lo que nos da

$$\kappa = \frac{\|V \wedge A\|}{v^3}.$$

Las dos últimas igualdades suponen que la imagen de x está en $\mathbb{R}^3$. En el lenguaje geométrico hemos obtenido las fórmulas siguientes:

$$T = \frac{x'}{\|x'\|}, \quad N = \frac{\|x'\|^2 x'' - (x'x'')x'}{\|x'\|\,\|x' \wedge x''\|}, \quad \kappa = \frac{\|x' \wedge x''\|}{\|x'\|^3.} \tag{4.4}$$

En el caso de curvas planas es conveniente modificar como sigue el vector normal y la curvatura. Fijada una orientación en el plano, definimos el vector normal de una curva $x(s)$ parametrizada por el arco como el vector unitario $N(s)$ que es perpendicular a $T(s)$ y de modo que la base $\big(T(s), N(s)\big)$ esté orientada positivamente. Esta definición puede diferir de la anterior en cuanto al signo de $N(s)$, por lo que redefinimos la curvatura de modo que $x''(s) = \kappa(s)N(s)$. Notemos que ahora el vector normal está definido incluso en los puntos donde la curvatura es nula.

Con el convenio usual de orientación, el vector N apunta hacia la izquierda si miramos en el sentido de T. La curvatura es positiva si cuando la curva avanza se desvía hacia la izquierda y negativa si se desvía hacia la derecha (o nula si no se desvía). Diremos que la curva gira en sentido positivo o en sentido negativo según el signo de su curvatura. El sentido de giro positivo es el contrario a las agujas del reloj. De este modo, la orientación distingue los dos sentidos de giro.

Vector binormal y torsión La teoría de curvas en $\mathbb{R}^3$ se completa con la introducción del vector binormal y la torsión. El vector *binormal* de una curva $x(s)$ tres veces derivable en un punto de curvatura no nula es $B(s) = T(s) \wedge N(s)$. La base formada por los vectores $\big(T(s), N(s), B(s)\big)$ se conoce como *triedro de Frenet.* Definimos la *torsión* de x en cada punto como $\tau(s) = -N'(s)B(s)$.

Para interpretar la torsión empezaremos por determinar N'. Digamos que

$$N' = aT + bN + cB. \tag{4.5}$$

Multiplicando por T obtenemos $a = N'T$, y como $TN = 0$, derivando resulta $T'N + TN' = 0$, luego $a = -T'N = -\kappa NN = -\kappa$.

Si multiplicamos (4.5) por N resulta $b = N'N$, pero al derivar en $NN = 1$ resulta $2NN' = 0$, luego $b = 0$. Por último, es claro que $c = N'B = -\tau$. Por consiguiente

$$N' = -\kappa T - \tau B.$$

La misma técnica nos da una expresión para B'. Sea $B' = aT + bN + cB$. Multiplicando por T obtenemos $a = TB'$, pero de $BT = 0$ se concluye que $TB' = -T'B = -\kappa NB = 0$. Similarmente $b = NB' = -N'B = \tau$. Al multiplicar por B llegamos a $c = B'B = 0$, pues $BB = 1$.

Tenemos así las llamadas *fórmulas de Frenet:*

$$\begin{aligned} T' &= \kappa N, \\ N' &= -\kappa T - \tau B, \\ B' &= \tau N. \end{aligned}$$

Vemos que si $\tau = 0$ en todo punto entonces B es constante, luego $(xB)' = x'B = TB = 0$ implica que xB es constante, luego x está contenido en un plano perpendicular a B, luego la curva es plana. El recíproco es claro. Así pues, las curvas sin torsión son exactamente las curvas planas. En general, el plano $x(s)+\langle T(s), N(s)\rangle$ recibe el nombre de *plano osculante* a la curva. Si la curva es plana, su plano osculante es el mismo en todo punto, y la curva está contenida en él. En caso contrario, puede probarse que el plano osculante en un punto es el plano más próximo a la curva en un entorno del punto. La tercera fórmula de Frenet muestra que la torsión mide la rapidez con que varía B o, lo que es lo mismo, la rapidez con la que varía el plano osculante.

A continuación derivamos una fórmula explícita para la torsión de una curva. Si x está parametrizada por la longitud de arco, entonces

$$N = \frac{x''}{\kappa}, \quad N' = \frac{x'''\kappa - x''\kappa'}{\kappa^2}.$$

luego,

$$\tau = -BN' = (T \wedge N)N' = -\frac{1}{\kappa}\,(x' \wedge x'')N' = -\frac{(x' \wedge x'')x'''}{\kappa^2} = -\frac{(x', x'', x''')}{\kappa^2},$$

donde $(u, v, w) = u(v \wedge w)$ es el *producto mixto* de vectores. Si la parametrización no es la natural tenemos evidentemente

$$B = \frac{x' \wedge x''}{\|x' \wedge x''\|},$$

y un cálculo rutinario nos lleva de la expresión que tenemos para $\tau(s)$ a

$$\tau = -\frac{(x', x'', x''')}{\|x' \wedge x''\|^2}.$$

Ejercicio: Calcular la curvatura y la torsión de la *hélice* $(r \operatorname{sen} t, r \cos t, kt)$.

Para acabar probaremos que la curvatura y la torsión determinan una curva salvo por su posición en el espacio.

Teorema 4.23 *Sean $x, \bar{x} : I \longrightarrow \mathbb{R}^3$ dos curvas con las mismas funciones κ y τ. Entonces existe una isometría $f : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ tal que $x(s) = f\big(\bar{x}(s)\big)$ para todo $s \in I$.*

Demostración: Es fácil comprobar que los vectores del triedro de Frenet, así como la curvatura y la torsión se conservan por isometrías, en el sentido de que, por ejemplo, si f es una isometría se cumple $T_{x\circ f}(s) = T_x \circ \vec{f}$, donde $\vec{f}$ es la isometría lineal asociada a f. Igualmente $\kappa_{x\circ f}(s) = \kappa(s)$, etc.

Sea $s_0 \in I$. Aplicando una isometría a $\bar{x}$ podemos exigir que $x(s_0) = \bar{x}(s_0)$, $T(s_0) = \bar{T}(s_0)$, $N(s_0) = \bar{N}(s_0)$, $B(s_0) = \bar{B}(s_0)$, $\kappa(s_0) = \bar{\kappa}(s_0)$, $\tau(s_0) = \bar{\tau}(s_0)$. Probaremos que en estas condiciones $x = \bar{x}$.

Es claro que

$$\frac{1}{2}\frac{d}{ds}\big(\|T-\bar{T}\|^2+\|N-\bar{N}\|^2+\|B-\bar{B}\|^2\big)$$
$$=(T-\bar{T})(T'-\bar{T}')+(N-\bar{N})(N'-\bar{N}')+(B-\bar{B})(B'-\bar{B}')$$

Aplicando las fórmulas de Frenet queda

$$\kappa(T-\bar{T})(N-\bar{N})-\kappa(N-\bar{N})(T-\bar{T})-\tau(N-\bar{N})(B-\bar{B})+\tau(B-\bar{B})(N-\bar{N})=0,$$

luego $T=\bar{T}$, $N=\bar{N}$, $B=\bar{B}$, pues las diferencias son constantes y se anulan en s_0. La primera igualdad es $x'=\bar{x}'$, y como ambas funciones coinciden en s_0 ha de ser $x=\bar{x}$. ■

Como consecuencia, si una curva plana tiene curvatura constante κ, entonces es un arco de circunferencia de radio $r=1/\kappa$, pues su curvatura y su torsión (nula) coinciden con las de la circunferencia. Las curvas de curvatura nula son las rectas.

Sistemas de referencia no inerciales A la hora de medir la posición de un objeto físico es necesario tomar a otro como referencia. Aunque en la práctica las coordenadas cartesianas no son siempre las más apropiadas, teóricamente podemos imaginar que seleccionamos un objeto rígido en el que marcamos cuatro puntos a los que asignamos las coordenadas $(0,0,0)$, $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ y con respecto a ellos determinamos la posición de cualquier otro punto del espacio.

Podría parecer conveniente exigir que el objeto mediante el cual definimos nuestro sistema de referencia esté en reposo, para evitar que las coordenadas de un cuerpo varíen por causa del movimiento del sistema de referencia y no por su propio movimiento. Sin embargo la Tierra, el Sol y las estrellas se mueven por el espacio, luego no tenemos a nuestra disposición ningún objeto del que podamos garantizar que está en reposo. Más aún, la física enseña que el requisito mismo no tiene sentido, pues no existe forma de dar sentido al concepto de "reposo absoluto", sino que el concepto de movimiento es esencialmente relativo al sistema de referencia que adoptemos.

Pese a ello, no todos los sistemas de referencia son equivalentes. Una clase especial de ellos la forman los determinados por objetos libres de toda influencia externa. Uno de los postulados básicos de la física es que si dos objetos están libres de toda influencia externa, su movimiento relativo será uniforme, es decir, al tomar como referencia a uno de ellos el otro se moverá (en línea recta) con velocidad constante, tal vez nula. A los sistemas de referencia en estas condiciones se les llama *inerciales,* de modo que la primera ley de Newton, de la que ya hemos hablado, afirma en realidad que la velocidad de un cuerpo libre de toda influencia exterior medida *desde un sistema de referencia inercial* permanece constante.

Esta ley no es aplicable a sistemas no inerciales. Por ejemplo, enseguida justificaremos que un tren que se mueve con velocidad constante puede ser

considerado como un sistema de referencia inercial, pero en el momento en que el tren acelera o frena deja de serlo. En efecto, imaginemos una esfera en reposo sobre el pasillo del tren. Visto desde la tierra, tanto el tren como la esfera están moviéndose a la misma velocidad. Cuando el tren frena, los frenos actúan sobre él haciéndolo parar, pero no actúan sobre la esfera, que sigue moviéndose a la misma velocidad. Desde un sistema de referencia determinado por el tren, la esfera en reposo pasa a moverse hacia delante sin que nada haya influido sobre ella. Esto es una violación de la primera ley de Newton.

La Tierra no es un sistema de referencia inercial a causa de sus movimientos de rotación y traslación. No obstante, la traslación alrededor del Sol sigue una órbita de radio tan grande que localmente es casi recta, y las consecuencias de este movimiento son inapreciables. En cuanto a la rotación, sus efectos sí son detectables mediante experimentos físicos que falsean la ley de Newton, pero en muchas ocasiones también resulta despreciable (una bola encima de una mesa nunca empieza a moverse sola como en el tren). Un sistema de referencia determinado por un objeto que se mueva a velocidad constante respecto a un sistema de referencia inercial cumple igualmente la ley de Newton y las leyes restantes de la dinámica, por lo que puede ser considerado inercial. Es el caso del tren que comentábamos antes. Lo que sucede es que aunque el tren está sometido a la acción de su motor, ésta se emplea únicamente en contrarrestar las fuerzas de rozamiento que se oponen a su avance, con lo que las dos acciones que éste experimenta se cancelan mutuamente, y el resultado es el mismo que si ninguna fuerza actuara sobre él.

Con más detalle: los cuerpos actúan unos sobre otros alterando su estado de movimiento. Por ejemplo, si dejamos en el aire un objeto en reposo éste no permanecerá en tal estado, sino que caerá, y esto no es una violación de la ley de Newton. Lo que sucede es que el cuerpo no está libre de acciones externas, sino que sufre la acción de la gravedad terrestre. La *segunda ley de Newton* afirma que la acción que un cuerpo ejerce sobre otro se traduce siempre en una aceleración sobre el mismo. Más detalladamente, a dicha acción se le puede asociar un vector llamado *fuerza,* de modo que si sumamos todas las fuerzas que produce sobre un cuerpo cada uno de los cuerpos externos que le influyen, el vector fuerza resultante es el producto de una constante, llamada *masa inercial* del cuerpo, por la aceleración que experimenta.

Ahora vamos a estudiar qué consecuencias tiene la rotación de la tierra en el movimiento de los objetos. En general, consideremos los vectores

$$i = (\cos \omega t, \operatorname{sen} \omega t, 0), \quad j = (-\operatorname{sen} \omega t, \cos \omega t, 0), \quad k = (0, 0, 1).$$

En cada instante t, estos vectores forman una base ortonormal positivamente orientada (notar que $k = i \wedge j$). Podemos suponer que hemos fijado un sistema de referencia inercial con origen en el centro de la tierra y que k apunta hacia el norte. Si ω es la velocidad angular de la Tierra, es decir, $\pi/12$ radianes por hora, entonces los tres vectores se mueven con la Tierra, y están en reposo respecto a ella.

Consideremos un objeto de masa m que se mueve según la trayectoria $x(t)$. Esto significa que la fuerza resultante que actúa sobre él en cada instante t es

$F = mx''$. Expresemos

$$x(t) = x_r(t)i(t) + y_r(t)j(t) + z_r(t)k.$$

Así, (x_r, y_r, z_r) son las coordenadas del móvil respecto a un sistema de referencia *relativo* a la Tierra. Derivemos:

$$x' = V_r - \omega y_r i + \omega x_r j = V_r + \omega k \wedge x,$$

donde $V_r = x_r' i + y_r' j + z_r' k$ es la *velocidad relativa* del móvil. Volvemos a derivar:

$$x'' = A_r + \omega k \wedge V_r + \omega k \wedge x' = A_r + 2\omega k \wedge V_r + \omega^2 k \wedge (k \wedge x),$$

donde $A_r = x_r'' i + y_r'' j + z_r'' k$ es la *aceleración relativa.* Multiplicamos por la masa m del móvil y despejamos

$$mA_r = F + 2m\omega V_r \wedge k + m\omega^2 (k \wedge x) \wedge k.$$

Ésta es la expresión de la segunda ley de Newton en un sistema de referencia en rotación con velocidad angular constante. La masa de un objeto por la aceleración que experimenta no es la resultante de las fuerzas que actúan sobre el objeto, sino la suma de esta resultante más dos "fuerzas ficticias", llamadas *fuerza centrífuga* y *fuerza de Coriolis,* dadas por

$$\begin{aligned} F_{\text{cen}} &= m\omega^2 (k \wedge x) \wedge k, \\ F_{\text{cor}} &= 2m\omega V_r \wedge k. \end{aligned}$$

Se las llama fuerzas ficticias (o inerciales) porque se comportan formalmente como fuerzas que hay que sumar a las demás, pero que no se corresponden con ninguna acción de ningún objeto sobre el móvil (como en el caso de la esfera que se mueve sola).

Si el móvil se encuentra en la superficie de la tierra a una latitud α (el ángulo que forma con el ecuador) entonces la fuerza centrífuga tiene la dirección y el sentido de x, es decir, apunta hacia arriba (recordemos que el origen de coordenadas está en el centro de la Tierra) y su módulo es $m\omega^2 R \cos\alpha$, donde R es el radio de la Tierra, luego es nula en los polos y máxima en el ecuador.

Para hacernos una idea de su intensidad, sobre un cuerpo de $1kg$ de masa actúa una fuerza centrífuga de

$$1 \cdot (7{,}29 \cdot 10^{-5})^2\, 6{,}3 \cdot 10^6 \approx 0{,}03N,$$

mientras[1] que su peso es de $9{,}8N$, unas 326 veces mayor. Por consiguiente la fuerza centrífuga es generalmente despreciable. Si la Tierra girara mucho más rápido los cuerpos pesarían menos por causa de esta fuerza y llegado a un punto podrían salir despedidos hacia el cielo.

La fuerza de Coriolis sólo actúa sobre cuerpos en movimiento respecto a la Tierra. Por ejemplo, si un cuerpo cae su velocidad V_r apunta al centro de la

[1] La unidad de fuerza es el Newton. Un Newton es la fuerza que aplicada a un cuerpo de $1Kg$ de masa le produce una aceleración de $1m/s^2$.

Tierra y $V_r \wedge k$ apunta hacia el este, por lo que los cuerpos que caen sufren una desviación hacia el este, nula en los polos y máxima en el ecuador.

Si el movimiento se realiza sobre la superficie de la Tierra observamos que k apunta hacia el exterior de la misma en el hemisferio norte y hacia el interior en el hemisferio sur, por lo que $V_r \wedge k$ apunta hacia la derecha de V_r en el hemisferio norte y hacia la izquierda en el hemisferio sur. Si descomponemos esta fuerza en dos vectores, uno en la dirección del centro de la Tierra y otro tangente a la misma, la gravedad hace inadvertible la primera componente, pero la segunda es apreciable. Ésta es nula en el ecuador y máxima en los polos. Puesto que se dirige siempre hacia el mismo lado, la fuerza de Coriolis hace girar los objetos, y se pone de manifiesto en los líquidos y gases, por ejemplo, el agua que cae por un desagüe gira en sentido horario en el hemisferio norte, en sentido antihorario en el hemisferio sur y no gira en las proximidades del ecuador. También puede apreciarse su efecto sobre un péndulo suficientemente largo y pesado (péndulo de Foucault).

***Longitud de arcos no euclídeos** Veamos ahora cómo se generalizan los resultados sobre longitud de arcos a las geometrías hiperbólica y elíptica. Comencemos por la primera. Consideremos un arco $X(t) = \big(x(t), y(t), z(t)\big)$ en el plano proyectivo $\mathrm{P}^2(\mathbb{R})$ contenido en el interior de una cónica de ecuación $f(X, X) = 0$. Llamamos $s(t)$ a la longitud de arco que queremos definir. Lo haremos determinando su derivada como en el caso euclídeo. La idea central es que si d representa la distancia hiperbólica entre dos puntos, entonces

$$\frac{s(t+h) - s(t)}{h} \approx \frac{d\big(X(t+h), X(t)\big)}{|h|},$$

y la aproximación será mejor cuanto menor sea h, pues si X es derivable se parece a una recta alrededor de $X(t)$, luego el límite cuando $h \to 0$ en el cociente de la izquierda nos dará la derivada de s. Para calcular este límite usaremos que

$$\lim_{v\to 0} \frac{\operatorname{senh} v}{v} = 1$$

y que

$$\operatorname{senh} d(X, Y) = \sqrt{\frac{f^2(X, Y) - f(X, X)f(Y, Y)}{f(X, X)f(Y, Y)}}.$$

Así pues,

$$\frac{ds}{dt} = \lim_{h\to 0} \frac{1}{|h|} \sqrt{\frac{f^2\big(X(t), X(t+h)\big) - f\big(X(t), X(t)\big)f\big(X(t+h), X(t+h)\big)}{f\big(X(t), X(t)\big)f\big(X(t+h), X(t+h)\big)}}.$$

Sea $X(t + h) = X(t) + \Delta X(t, h)$. Usando que f es bilineal la expresión anterior se simplifica:

$$\frac{ds}{dt} = \lim_{h\to 0} \frac{1}{|h|} \sqrt{\frac{f^2(X, \Delta X) - f(X, X)f(\Delta X, \Delta X)}{f(X, X)\big(f(X, X) + 2f(X, \Delta X) + f(\Delta X, \Delta X)\big)}}$$

$$= \lim_{h\to 0}\sqrt{\frac{f^2(X,\frac{\Delta X}{h}) - f(X,X)f(\frac{\Delta X}{h},\frac{\Delta X}{h})}{f(X,X)\big(f(X,X)+2f(X,\Delta X)+f(\Delta X,\Delta X)\big)}}$$

$$= \sqrt{\frac{f^2(X,X') - f(X,X)f(X',X')}{f^2(X,X)}}.$$

Vamos a particularizar esta fórmula para el caso del plano de Klein, es decir, tomando como f la circunferencia unidad $f(X_1,X_2) = z_1z_2 - x_1x_2 - y_1y_2$ y el arco de la forma $X(t) = \big(x(t),y(t),1\big)$, de modo que $X'(t) = \big(x'(t),y'(t),0\big)$. El resultado es:

$$\left(\frac{ds}{dt}\right)^2 = \frac{(xx'-yy')^2 + (1-x^2-y^2)(x'^2+y'^2)}{(1-x^2-y^2)^2}.$$

Operando y multiplicando por dt^2 obtenemos una expresión para el elemento de longitud hiperbólica en el plano de Klein:

$$ds^2 = \frac{dx^2+dy^2-(x\,dy-y\,dx)^2}{(1-x^2-y^2)^2},$$

que ha de entenderse como una ecuación funcional, para cada punto t, entre la diferencial de la longitud de arco $ds(t)$ y las diferenciales $dx(t)$, $dy(t)$ de las funciones coordenadas del arco. Formalmente, si $X:[a,b]\longrightarrow \mathbb{R}^2$, definimos $s(t)$ como la integral desde a hasta t de la raíz cuadrada del miembro derecho de la igualdad anterior, con lo que obtenemos una función que satisface dicha relación. Informalmente ds así calculado es el incremento infinitesimal que experimenta la longitud de arco cuando el parámetro se incrementa en dt y la integral de ds nos da el incremento completo de la longitud de arco entre dos límites dados.

Observemos que la longitud hiperbólica es invariante por isometrías. En efecto, a cada arco $X:[a,b]\longrightarrow \mathbb{R}^2$ le hemos asociado la función s determinada por $s(a)=0$ y

$$\frac{ds}{dt} = \lim_{h\to 0}\frac{d\big(X(t+h),X(t)\big)}{|h|}$$

y es obvio que el miembro derecho es invariante por isometrías.

La expresión que hemos obtenido no es muy manejable, y las integrales a que da lugar resultan complicadas. Las coordenadas polares nos dan una expresión más sencilla. Si diferenciamos $(x,y) = (r\cos\theta, r\operatorname{sen}\theta)$ y sustituimos en la expresión de ds obtenemos

$$ds^2 = \frac{dr^2}{(1-r^2)^2} + \frac{r^2}{1-r^2}\,d\theta^2.$$

Si un punto se encuentra a una distancia hiperbólica ρ del centro del plano de Klein, la distancia euclídea es $r = \tanh\rho$. Al diferenciar esta relación y

sustituir en la expresión anterior obtenemos el elemento de longitud hiperbólico en coordenadas polares hiperbólicas, que resulta ser

$$ds^2 = d\rho^2 + \operatorname{senh}^2 \rho \, d\theta^2, \tag{4.6}$$

relación análoga a (4.3).

Por ejemplo, consideremos una circunferencia cuyo centro coincida con el del plano de Klein y de radio (hiperbólico) r. Entonces su longitud hiperbólica es

$$\int_0^{2\pi} \operatorname{senh} r \, d\theta = 2\pi \operatorname{senh} r.$$

Observemos que si el radio es pequeño la longitud es aproximadamente la euclídea $2\pi r$.

La fórmula (4.6) es intrínseca, en el sentido de que las coordenadas (ρ, θ) de un punto P representan la distancia hiperbólica de P a un punto fijo O y el ángulo de la semirrecta $\overrightarrow{OP}$ con una semirrecta fija de origen O, y nada de esto depende del plano de Klein. Por lo tanto la fórmula ha de ser válida también en el círculo de Poincaré. La relación entre la distancia euclídea r y la distancia hiperbólica ρ de un punto P al punto 0 en el círculo de Poincaré es

$$\rho = \log \frac{1+r}{1-r},$$

de donde

$$\operatorname{senh} \rho = \frac{2r}{1-r^2}, \quad d\rho = \frac{2dr}{1-r^2}.$$

Por consiguiente, el elemento de longitud en coordenadas polares (euclídeas) en el círculo de Poincaré resulta ser

$$ds = 2\,\frac{\sqrt{dr^2 + r^2\, d\theta^2}}{1-r^2}.$$

Según (4.3), el numerador es el elemento de longitud euclídea en coordenadas polares. Si usamos la notación compleja para el arco $z(t) = x(t) + i\,y(t)$ y llamamos $|dz| = \sqrt{dx^2 + dy^2}$ al elemento de longitud euclídea, entonces tenemos

$$ds = \frac{2\,|dz|}{1-|z|^2}.$$

Esta expresión diferencial muestra la naturaleza de la distancia hiperbólica mucho más claramente que la fórmula de la distancia entre dos puntos. Vemos que la distancia hiperbólica es infinitesimalmente la euclídea dividida entre el factor más simple posible que hace que se "dilate" al acercarnos al borde del círculo de radio 1, de modo que las pequeñas distancias euclídeas son cada vez más grandes desde el punto de vista hiperbólico. La expresión también es muy clara en el semiplano de Poincaré. La transformación circular

$$z = \frac{iw+1}{w+i}$$

convierte el círculo $|z| < 1$ en el semiplano $\operatorname{Im} w > 0$. Se comprueba[2] que

$$dz = \frac{-2\,dw}{(w+i)^2}, \quad |dz| = \frac{2\,|dw|}{|w+i|^2}, \quad 1 - |z|^2 = \frac{4y}{|w+i|^2}.$$

De todo esto resulta

$$ds = \frac{|dw|}{y},$$

es decir, que el elemento de longitud hiperbólico en el semiplano de Poincaré es el euclídeo dividido entre la parte imaginaria, de modo que cuando ésta tiende a 0 las longitudes tienden a infinito.

Todos los argumentos anteriores valen igualmente para la geometría elíptica, partiendo ahora de una cónica imaginaria $f(X_1, X_2)$. La única diferencia es un signo en la fórmula

$$\operatorname{sen} d(X,Y) = \sqrt{\frac{-f^2(X,Y) + f(X,X)f(Y,Y)}{f(X,X)f(Y,Y)}},$$

debido a que la relación fundamental entre los senos y cosenos circulares difiere en un signo respecto a la hiperbólica. Como consecuencia llegamos a

$$\left(\frac{ds}{dt}\right)^2 = \frac{-f^2(X,X') + f(X,X)f(X',X')}{f^2(X,X)}.$$

En el modelo esférico, o sea, si suponemos $f(X_1, X_2) = x_1x_2 + y_1y_2 + z_1z_2$ y $f(X,X) = 1$, al derivar respecto de t queda $f(X,X') = 0$, y la expresión se reduce a

$$\left(\frac{ds}{dt}\right)^2 = f(X',X') = x'^2 + y'^2 + z'^2,$$

es decir,

$$ds^2 = dx^2 + dy^2 + dz^2,$$

luego la longitud elíptica de un arco contenido en la esfera unitaria coincide con su longitud euclídea. En coordenadas polares elípticas:

$$(x,y,z) = (\operatorname{sen}\rho\cos\theta, \cos\rho\operatorname{sen}\theta, \operatorname{sen}\theta),$$

la expresión es

$$ds^2 = d\rho^2 + \operatorname{sen}^2\rho\, d\theta.$$

La longitud de una circunferencia elíptica de radio (elíptico) r es $2\pi \operatorname{sen} r$.

[2]La teoría de funciones de variable compleja justifica que es lícito calcular dz derivando la expresión anterior como si w fuera una variable real y tratando las constantes imaginarias como constantes reales. Nosotros no hemos probado esto, por lo que el lector puede, si lo desea, hacer los cálculos en términos de las dos variables reales x, y, pero está avisado de que llegará al mismo resultado.

Capítulo V

Introducción a las variedades diferenciables

En este capítulo aplicaremos el cálculo diferencial al estudio de las superficies. Si bien todos los ejemplos que consideraremos serán bidimensionales, la mayor parte de la teoría la desarrollaremos sobre un concepto general de "superficie de n dimensiones". La idea básica es que una superficie es un espacio topológico que localmente se parece a un plano. El ejemplo típico es la superficie terrestre: tenemos que alejarnos mucho de ella para darnos cuenta de que no es plana, sino esférica. Una definición topológica que recoja estas ideas sería la siguiente:

> Un subconjunto S de $\mathbb{R}^n$ es una superficie si para cada punto $p \in S$ existe un entorno V de p, un abierto U en $\mathbb{R}^2$ y un homeomorfismo $X : U \longrightarrow V \cap S$.

Es decir, S es una superficie si alrededor de cada punto es homeomorfa a un abierto de $\mathbb{R}^2$. Notar que no pedimos que S sea homeomorfa a un abierto de $\mathbb{R}^2$, sino sólo que lo sea alrededor de cada punto. Basta pensar en una esfera para comprender la importancia de este hecho. Una esfera no es homeomorfa a un abierto de $\mathbb{R}^2$, pero un pequeño trozo de esfera es como un trozo de plano abombado, homeomorfo a un trozo de plano "llano". Sin embargo nosotros estamos interesados en superficies diferenciables, en el sentido de que se parezcan a planos afines alrededor de cada punto. Podría pensarse que para conseguir esto bastaría exigir que el homeomorfismo X sea diferenciable, pero no es así. Por ejemplo, pensemos en $X(u, v) = \big(u^3, v, |u^3|\big)$. La aplicación X es diferenciable, y es un homeomorfismo entre $\mathbb{R}^2$ y un conjunto $S \subset \mathbb{R}^3$ cuya forma es la de una hoja de papel doblada por la mitad. Alrededor de los puntos de la forma $(0, v, 0)$ no se parece a ningún plano, sino que tiene un "pico". Si la Tierra tuviera esta forma no necesitaríamos alejarnos de ella para darnos cuenta de que estaría "doblada".

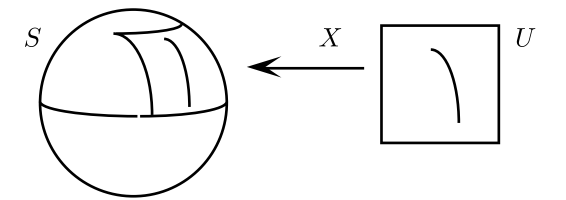
La razón es que $dX(0,b) = (0, dv(0,b), 0)$, de modo que alrededor de un punto $(0,b)$ la función X se parece a la aplicación afín $f(u,v) = (0,v,0)$, cuya imagen es la recta $x = z = 0$. Así pues, aunque topológicamente la imagen de X es homeomorfa a un plano, desde el punto de vista del cálculo diferencial la imagen de X alrededor de un punto $(0,y,0)$ se parece a la recta $x = z = 0$, y no a un plano. Para evitar esto hemos de exigir que la imagen de $dX(u,v)$ sea un plano y no una recta. Esto es tanto como decir que la matriz jacobiana tenga rango 2.

5.1 Variedades

Definición 5.1 Un conjunto $S \subset \mathbb{R}^m$ es una *variedad diferenciable* de dimensión $n \leq m$ y de clase C^q si para cada punto $p \in S$ existe un entorno V de p, un abierto U en $\mathbb{R}^n$ y una función $X : U \longrightarrow \mathbb{R}^m$ de clase C^q de modo que el rango de la matriz JX sea igual a n en todo punto y $X : U \longrightarrow S \cap V$ sea un homeomorfismo. Una aplicación X en estas condiciones se llama *carta* de S alrededor de p.

En lo sucesivo supondremos que las variedades con las que trabajamos son de clase C^q para un q suficientemente grande como para que existan las derivadas que consideremos (y sean continuas). Rara vez nos hará falta suponer $q > 3$, aunque de hecho todos los ejemplos que consideraremos serán de clase C^∞.

La palabra "carta" hay que entenderla en el sentido de "mapa". En efecto, podemos pensar en U como un mapa "plano" de una región de S, y la aplicación X es la que hace corresponder cada punto del mapa con el punto real que representa.



Alternativamente, podemos pensar en X^{-1} como una aplicación que asigna a cada punto $p \in S \cap V$ unas *coordenadas* $x = (x_1, \ldots, x_n) \in U \subset \mathbb{R}^n$, de forma análoga a los sistemas de coordenadas en un espacio afín.[1] Dentro de poco será equivalente trabajar con cartas o con sistemas de coordenadas, pero por el momento podemos decir que las cartas son diferenciables y en cambio no tiene sentido decir que las funciones coordenadas lo sean, pues no están definidas sobre abiertos de $\mathbb{R}^m$.

[1] Etimológicamente, una "variedad" no es más que un conjunto cuyos elementos vienen determinados por "varias" coordenadas. En los resultados generales llamaremos $x_1, \ldots, x_n$ a las coordenadas para marcar la analogía con $\mathbb{R}^n$, aunque en el caso de curvas seguiremos usando la variable t (o s si la parametrización es la natural) y en el caso de superficies $S \subset \mathbb{R}^3$ usaremos x, y, z para las coordenadas en $\mathbb{R}^3$ y u, v para las coordenadas en S.

Ejemplo Todo abierto U en $\mathbb{R}^n$ es una variedad diferenciable de dimensión n y clase C^∞. Basta tomar como carta la identidad en U. De este modo, todos los resultados sobre variedades valen en particular para $\mathbb{R}^n$ y sus abiertos. ■

Ejercicio: Refinar el argumento del teorema 2.21 para concluir que dos puntos cualesquiera de una variedad conexa S de clase C^q pueden ser unidos por un arco de clase C^q contenido en S.

El teorema siguiente proporciona una clase importante de variedades diferenciables, pues a continuación vemos que toda variedad es localmente de este tipo.

Teorema 5.2 *Sea $f : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}^k$ una aplicación de clase C^q sobre un abierto U y $X : U \longrightarrow \mathbb{R}^{n+k}$ la aplicación dada por $X(x) = \big(x, f(x)\big)$. Entonces $X[U]$ es una variedad diferenciable de dimensión n y clase C^q.*

Demostración: Basta observar que X es obviamente un homeomorfismo en su imagen (su inversa es una proyección) y $JX(x)$ contiene una submatriz de orden n igual a la identidad, luego su rango es n. La definición se satisface tomando $V = \mathbb{R}^{n+k}$. ■

Observar que $X[U]$ es la gráfica de f, luego el teorema anterior afirma que la gráfica de una función diferenciable es siempre una variedad diferenciable. Ahora veamos que todo punto de una variedad diferenciable tiene un entorno en el que la variedad es la gráfica de una función.

Teorema 5.3 *Sea $S \subset \mathbb{R}^{n+k}$ una variedad de clase C^q y de dimensión n. Sea $p \in S$. Entonces existe un entorno V de p, un abierto U en $\mathbb{R}^n$ y una función $f : U \longrightarrow \mathbb{R}^k$ de clase C^q de modo que la aplicación $X : U \longrightarrow \mathbb{R}^{n+k}$ dada por $X(x) = \big(x, f(x)\big)$ es una carta alrededor de p.*

En realidad hemos de entender que las coordenadas de x y $f(x)$ se intercalan en un cierto orden que no podemos elegir, tal y como muestra la prueba.

Demostración: Sea $Y : W \longrightarrow \mathbb{R}^{n+k}$ una carta alrededor de p. Sea V un entorno de p tal que $Y : W \longrightarrow S \cap V$ sea un homeomorfismo. Sea $t_0 \in W$ el vector de coordenadas de p, es decir, $Y(t_0) = p$. Puesto que $JY(t_0)$ tiene rango n, reordenando las funciones coordenadas de Y podemos suponer que el determinante formado por las derivadas parciales de las n primeras es no nulo. Digamos que $Y(t) = \big(Y_1(t), Y_2(t)\big)$, donde $|JY_1(t_0)| \neq 0$. Sea $p_1 = Y_1(t_0)$.

Por el teorema de inyectividad local y el teorema de la función inversa, existe un entorno abierto $G \subset W$ de t_0 tal que Y_1 es inyectiva en G, $Y_1[G] = U$ es abierto en $\mathbb{R}^n$ y la función $Y_1^{-1} : U \longrightarrow G$ es de clase C^q.

El conjunto $Y[G]$ es un entorno abierto de p en $S \cap V$, luego existe un entorno abierto V' de p en $\mathbb{R}^{n+k}$ tal que $Y[G] = S \cap V \cap V'$. Cambiando V' por $V \cap V'$ podemos suponer que $V' \subset V$ y que $Y[G] = S \cap V'$.

De este modo, cada punto $p \in S \cap V'$ está determinado por sus coordenadas $t \in G$, las cuales a su vez están determinadas por $x = Y_1(t) \in U$, con la particularidad de que x es el vector de las primeras componentes de p. Concretamente

p está formado por x y $f(x) = Y_2\big(Y_1^{-1}(x)\big)$. La función f es de clase C^q en U. De este modo, si $x \in U$ y $t = Y_1^{-1}(x) \in G$, tenemos que

$$X(x) = \big(x, f(x)\big) = \big(Y_1(t), Y_2(t)\big) = Y(t) \in S \cap V'.$$

Recíprocamente, si $(x, y) \in S \cap V'$, entonces $x = Y_1(t)$, $y = Y_2(t)$ para un cierto $t \in G$, luego $(x, y) = \big(x, f(x)\big) = X(x)$. En definitiva tenemos que $X : U \longrightarrow S \cap V'$ es biyectiva. Su inversa es la proyección en las primeras componentes, luego X es un homeomorfismo. ■

En las condiciones de la prueba anterior, sea $\pi : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ la proyección en las n primeras componentes y $g : V' \longrightarrow G$ la aplicación dada por $g(x) = Y_1^{-1}\big(\pi(x)\big)$. Notemos que si $x \in V'$ entonces $\pi(x) \in U$, luego g está bien definida y es de clase C^q. Si $t \in G$ entonces

$$g\big(Y(t)\big) = g\big(Y_1(t), Y_2(t)\big) = Y_1^{-1}\big(Y_1(t)\big) = t,$$

luego $(Y|_G)^{-1}$ es la restricción a $V' \cap S$ de g. Con esto hemos probado:

Teorema 5.4 *Sea $Y : U \longrightarrow S \subset \mathbb{R}^m$ una carta de una variedad diferenciable de dimensión n y clase C^q. Para cada punto $t \in U$ existe un entorno $G \subset U$ de t, un entorno V de $Y(t)$ y una aplicación $g : V \longrightarrow G$ de clase C^q tal que $(Y|_G)^{-1} = g|_{V \cap S}$.*

De aquí se sigue una propiedad fundamental de las cartas:

Teorema 5.5 *Sea $S \subset \mathbb{R}^m$ una variedad diferenciable de dimensión n y de clase C^q. Sea $p \in S$ y $X : U \longrightarrow S \cap V$, $Y : U' \longrightarrow S \cap V'$ dos cartas alrededor de p. Sean $V_0 = V \cap V'$, $U_0 = X^{-1}[V_0]$, $U_0' = Y^{-1}[V_0]$. Entonces la aplicación $X \circ Y^{-1} : U_0 \longrightarrow U_0'$ es biyectiva, de clase C^q y con determinante jacobiano no nulo, con lo que su inversa es también de clase C^q.*

Demostración: Si $t \in U_0$, por el teorema anterior existe una función g de clase C^q definida en un entorno de $X(t)$ de modo que $X \circ Y^{-1} = X \circ g$ (en un entorno de t), luego $X \circ Y^{-1}$ es de clase C^q en un entorno de t, luego en todo U_0. Lo mismo vale para su inversa $Y \circ X^{-1}$, luego la regla de la cadena nos da que sus diferenciales son mutuamente inversas, luego los determinantes jacobianos son no nulos. ■

Veremos ahora otro ejemplo importante de variedades diferenciables. Primeramente consideraremos el caso lineal al cual generaliza.

Ejemplo Una variedad afín de dimensión n en $\mathbb{R}^m$ es también una variedad diferenciable de la misma dimensión y de clase C^∞. En efecto, una tal variedad está formada por los puntos que satisfacen un sistema de $m - n$ ecuaciones lineales linealmente independientes. Esto implica que la matriz de coeficientes del sistema tiene un determinante de orden $m - n$ no nulo, luego agrupando adecuadamente las variables podemos expresar el sistema como $xA + yB = c$, donde $x \in \mathbb{R}^n$, $y \in \mathbb{R}^{m-n}$, $|B| \neq 0$, luego podemos despejar $y = f(x) =$

$(c-xA)B^{-1}$, donde la función f es obviamente de clase C^∞. Esto significa que la variedad lineal está formada por los puntos (x, y) tales que $y = f(x)$, luego es la gráfica de f y por consiguiente es una variedad de clase C^∞. ■

Ahora probamos que las soluciones de un sistema de $k = m-n$ ecuaciones diferenciables con m incógnitas constituyen una variedad de dimensión n supuesto que se cumpla una condición de independencia similar a la independencia lineal que exigíamos en el ejemplo anterior.

Definición 5.6 Si $f : A \subset \mathbb{R}^{n+k} \longrightarrow \mathbb{R}^k$ es diferenciable en $(x, y) \in A$, donde $x \in \mathbb{R}^n$, $y \in \mathbb{R}^k$, definimos

$$\frac{\partial(f_1 \cdots f_k)}{\partial(y_1 \cdots y_k)}(x, y) = \begin{vmatrix} D_{n+1}f_1(x, y) & \cdots & D_{n+1}f_k(x, y) \\ \vdots & & \vdots \\ D_{n+k}f_1(x, y) & \cdots & D_{n+k}f_k(x, y) \end{vmatrix}.$$

Teorema 5.7 (Teorema de la función implícita) *Consideremos una aplicación $f : A \subset \mathbb{R}^{n+k} \longrightarrow \mathbb{R}^k$ de clase C^q en el abierto A, con $q \geq 1$. Sea $(x^0, y^0) \in A$ tal que $f(x^0, y^0) = 0$ y supongamos que*

$$\frac{\partial(f_1 \cdots f_k)}{\partial(y_1 \cdots y_k)}(x^0, y^0) \neq 0.$$

Entonces existen abiertos $V \subset A$, $U \subset \mathbb{R}^n$ de modo que $(x^0, y^0) \in A$, $x^0 \in U$ y una función $g : U \longrightarrow \mathbb{R}^k$ de clase C^q tal que

$$\{(x, y) \in V \mid f(x, y) = 0\} = \{(x, y) \in \mathbb{R}^{n+k} \mid x \in U,\ y = g(x)\}.$$

DEMOSTRACIÓN: Sea $F : A \longrightarrow \mathbb{R}^{n+k}$ la función $F(x, y) = (x, f(x, y))$. Sus funciones coordenadas son las proyecciones en las componentes de $\mathbb{R}^n$ más las funciones coordenadas de f, luego F es de clase C^q. Su determinante jacobiano es

$$\begin{vmatrix} 1 & \cdots & 0 & D_1f_1(x, y) & \cdots & D_1f_k(x, y) \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 1 & D_nf_1(x, y) & \cdots & D_nf_k(x, y) \\ 0 & \cdots & 0 & D_{n+1}f_1(x, y) & \cdots & D_{n+1}f_k(x, y) \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \cdots & 0 & D_{n+k}f_1(x, y) & \cdots & D_{n+k}f_k(x, y) \end{vmatrix},$$

que claramente coincide (salvo signo) con

$$\frac{\partial(f_1 \cdots f_k)}{\partial(y_1 \cdots y_k)}(x, y).$$

Así $|J_F(x^0, y^0)| \neq 0$. Por el teorema de inyectividad local existe un entorno V de (x^0, y^0) donde F es inyectiva y su determinante jacobiano no se anula. Por

el teorema de la función inversa tenemos que $W = F[V]$ es abierto en $\mathbb{R}^{n+k}$ y la función $G = F^{-1} : W \longrightarrow V$ es de clase C^q.

Podemos expresar $G(x, y) = \big(G_1(x, y), G_2(x, y)\big)$. Claramente G_1 y G_2 son ambas de clase C^q. Si $(x, y) \in W$, entonces

$$(x, y) = F\big(G(x, y)\big) = F\big(G_1(x, y), G_2(x, y)\big) = \Big(G_1(x, y), f\big(G(x, y)\big)\Big),$$

luego $G_1(x, y) = x$, con lo que en general $G(x, y) = \big(x, G_2(x, y)\big)$.

Definimos $U = \{x \in \mathbb{R}^n \mid (x, 0) \in W\}$. Es claro que se trata de un abierto. Además, $F(x^0, y^0) = (x^0, 0) \in W$, luego $x^0 \in U$. Definimos $g : U \longrightarrow \mathbb{R}^k$ mediante $g(x) = G_2(x, 0)$. Claramente g es de clase C^q.

Tomemos ahora $x \in U$ e $y = g(x)$. Hemos de probar que $(x, y) \in V$ y $f(x, y) = 0$. En efecto, por definición de U es $(x, 0) \in W$, luego $G(x, 0) \in V$, pero

$$G(x, 0) = \big(x, G_2(x, 0)\big) = \big(x, g(x)\big) = (x, y).$$

Además $(x, 0) = F\big(G(x, 0)\big) = F(x, y) = \big(x, f(x, y)\big)$, luego $f(x, y) = 0$.

Recíprocamente, si $(x, y) \in V$ y $f(x, y) = 0$ entonces

$$F(x, y) = \big(x, f(x, y)\big) = (x, 0) \in W,$$

luego $x \in U$ y $(x, y) = G\big(F(x, y)\big) = G(x, 0) = \big(x, G_2(x, 0)\big) = \big(x, g(x)\big)$, con lo que $g(x) = y$. ■

Lo que afirma este teorema es que si $S = \{x \in \mathbb{R}^{n+k} \mid f(x) = 0\}$ es un conjunto determinado por un sistema de k ecuaciones de clase C^q y $p \in S$ cumple la hipótesis entonces $V \cap S = X[U]$, donde $X(x) = (x, g(x))$, de donde se sigue que X cumple las condiciones para ser una carta de S alrededor de p. Si la hipótesis se cumple en todo punto entonces S es una variedad diferenciable de dimensión n.

Por ejemplo, si $f(x, y, z) = x^2 + y^2 + z^2 - r^2$, entonces el conjunto S es una esfera. Para comprobar que se trata de una superficie de clase C^∞ basta comprobar que en cada punto al menos una de las derivadas

$$\frac{\partial f}{\partial x} = 2x, \quad \frac{\partial f}{\partial y} = 2y, \quad \frac{\partial f}{\partial z} = 2z,$$

es no nula, pero las tres sólo se anulan simultáneamente en $(0, 0, 0)$, que no es un punto de S, luego, efectivamente, la esfera es una superficie diferenciable.

Es importante observar que la derivada que no se anula no siempre es la misma. Por ejemplo, en el polo norte $(0, 0, r)$ la única derivada que no se anula es la de z, luego en un entorno podemos expresar z como función $z(x, y)$. Concretamente, $z = \sqrt{r^2 - x^2 - y^2}$. Similarmente, la porción de esfera alrededor del polo sur es la gráfica de la función $z = -\sqrt{r^2 - x^2 - y^2}$. En cambio, alrededor de $(r, 0, 0)$ la esfera no es la gráfica de ninguna función $z(x, y)$. Es fácil ver que dado cualquier entorno U de $(r, 0, 0)$ y cualquier entorno V de $(r, 0)$ siempre hay puntos (x, y) en U para los cuales hay dos puntos distintos $(x, y, \pm z)$ en U

(con lo que (x, y) debería tener dos imágenes) y puntos (x, y) con $x^2 + y^2 > r^2$ para los que no existe ningún z tal que $(x, y, z) \in U$. Sin embargo, alrededor de este punto la esfera es la gráfica de la función $x = \sqrt{r^2 - y^2 - z^2}$.

El mismo argumento prueba en general que la *esfera* de dimensión n

$$S^n = \{x \in \mathbb{R}^{n+1} \mid \|x\|_2^2 = 1\}$$

es una variedad diferenciable.

Ejemplo: superficies de revolución Sea C una variedad diferenciable de dimensión 1 en $\mathbb{R}^2$. Supongamos que todos sus puntos (x, z) cumplen $x > 0$. Llamaremos *superficie de revolución* generada por C al conjunto

$$S = \big\{(x, y, z) \in \mathbb{R}^3 \mid \big(\sqrt{x^2 + y^2}, z\big) \in C\big\}.$$

El conjunto S está formado por todos los puntos que resultan de girar alrededor del eje Z los puntos de C. Vamos a ver que se trata de una variedad diferenciable de dimensión 2.

Tomemos $(x_0, y_0, z_0) \in S$ y $\bar{x}_0 = \sqrt{x_0^2 + y_0^2}$. Entonces $(\bar{x}_0, z_0) \in C$. Sea $\alpha(u) = \big(r(u), z(u)\big)$ una carta de C alrededor de este punto, digamos $r(u_0) = \bar{x}_0$, $z(u_0) = z_0$. Por definición existe un entorno V_0 de u_0 y un entorno U_0 de $(\bar{x}_0, z_0)$ de modo que $C \cap U_0 = \alpha[V_0]$. Sea

$$X(u, v) = \big(r(u) \cos v, r(u) \operatorname{sen} v, z(u)\big), \quad (u, v) \in V_0 \times \mathbb{R}.$$

Claramente X es diferenciable (de la misma clase que α) y su matriz jacobiana es

$$JX(u, v) = \begin{pmatrix} r'(u) \cos v & r'(u) \operatorname{sen} v & z'(u) \\ -r(u) \operatorname{sen} v & r(u) \cos v & 0 \end{pmatrix}.$$

El menor formado por las dos primeras columnas es $r(u)r'(u)$. Por hipótesis r no se anula y, por ser α una carta, su matriz jacobiana (r', z') no puede ser nula tampoco, luego si $r'(u) = 0$, entonces $z'(u) \neq 0$, luego uno de los menores $r(u)z'(u) \operatorname{sen} v$ o $-r(u)z'(u) \cos v$ es no nulo. En cualquier caso el rango de JX es 2.

Es claro que existe un $v_0 \in \mathbb{R}$ tal que $X(u_0, v_0) = (x_0, y_0, z_0)$. La aplicación X no es inyectiva, pero sí lo es su restricción a $V = V_0 \times]v_0 - \pi, v_0 + \pi[$. Veamos que es una carta para el punto dado. Sea

$$U = \{(x \cos v, x \operatorname{sen} v, z) \mid (x, z) \in U_0, |v - v_0| < \pi\}.$$

Sin la restricción sobre v, el conjunto U sería la antiimagen de U_0 por la aplicación continua $(x, y, z) \mapsto \big(\sqrt{x^2 + y^2}, z\big)$. En realidad U es la intersección de este abierto con el complementario del semiplano formado por los puntos $(x \cos v_0, x \operatorname{sen} v_0, z)$, con $x \geq 0$, que es un cerrado, luego U es abierto. Es fácil ver que $X[V] = U \cap S$. Falta probar que X^{-1} es continua, ahora bien, dado $(x, y, z) \in U \cap S$ podemos obtener su coordenada u como $u = \alpha^{-1}(\sqrt{x^2 + y^2}, z)$, que es una aplicación continua, y su coordenada v se obtiene aplicando a

$\big(x/r(u), y/r(u)\big)$ la inversa del homeomorfismo $v \mapsto (\cos v, \operatorname{sen} v)$ definido para $|v - v_0| < \pi$. Por lo tanto $X|_V$ es una carta alrededor de (x_0, y_0, z_0).

Si la variedad C es cubrible por una única carta $\big(r(u), z(u)\big)$, lo que se traduce en que C es una curva regular, entonces tenemos una única función X, de modo que todo punto de S admite como carta a una restricción de X. Expresaremos esto diciendo simplemente que X es una carta de S.

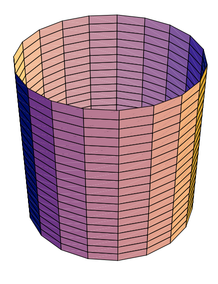
Las líneas de la forma $X(u, v_0)$ y $X(u_0, v)$, donde u_0 y v_0 son constantes, se llaman *meridianos* y *paralelos* de la superficie S. Los paralelos son siempre circunferencias paralelas entre sí, los meridianos son giros de la curva C.

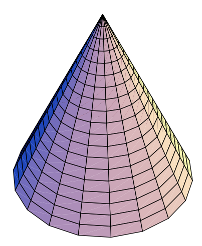
Los ejemplos más simples de superficies de revolución se obtienen al girar una recta. Si ésta es paralela al eje de giro obtenemos un *cilindro,* y en caso contrario un *cono.* En el caso del cono hemos de considerar en realidad una semirrecta abierta $\big(r(u), z(u)\big) = (mu, u)$, para $u > 0$, pues para $u = 0$ tenemos el vértice del cono, donde S no es diferenciable. Una carta del cilindro es

$$X(u, v) = (r\cos v, r\operatorname{sen} v, u),$$

y para el cono tenemos

$$X(u, v) = (mu\cos v, mu\operatorname{sen} v, u).$$





Las figuras muestran algunos de los meridianos y paralelos del cilindro y el cono. Los meridianos son rectas y los paralelos circunferencias.

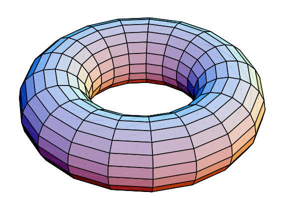
Un caso más sofisticado aparece al girar una circunferencia de radio r cuyo centro esté a una distancia R del eje Z, es decir, tomando

$$\big(r(u), z(u)\big) = (R + r\cos u, r\operatorname{sen} u), \quad \text{con } 0 < r < R.$$

Todo punto de la circunferencia admite como carta a una restricción de esta curva. Así obtenemos un tubo de sección circular cerrado sobre sí mismo. Recibe el nombre de *toro.*[2] En este caso

$$X(u, v) = (R\cos v + r\cos u\cos v, R\operatorname{sen} v + r\cos u\operatorname{sen} v, r\operatorname{sen} u).$$

[2]Del latín *torus,* que es el nombre dado en arquitectura a los salientes tubulares de las columnas. Obviamente no tiene nada que ver con *taurus,* el animal del mismo nombre en castellano.



Obviamente X es de clase C^∞. Su restricción a $]0,2\pi[\times]0,2\pi[$ es inyectiva y cubre todos los puntos del toro excepto los de las circunferencias $u=0$ y $v=0$. Si llamamos U al complementario de la unión de estas dos circunferencias tenemos un abierto en $\mathbb{R}^3$, y es claro que con él se cumple la definición de variedad. Igualmente se prueba que la restricción a $]-\pi,\pi[\times]-\pi,\pi[$ constituye una carta para los puntos exceptuados. Así pues, el toro es una superficie diferenciable de clase C^∞. Sus meridianos son circunferencias de radio r.

La esfera menos dos puntos antípodas puede considerarse como la superficie de revolución generada por la semicircunferencia $(r\operatorname{sen}\phi, r\cos\phi)$, para $\phi\in]0,\pi[$. La carta correspondiente es

$$X(\phi,\theta)=(r\operatorname{sen}\phi\cos\theta, r\operatorname{sen}\phi\operatorname{sen}\theta, r\cos\phi), \quad \phi\in]0,\pi[\,,\theta\in]0,2\pi[\,.$$

Si $p=X(\phi,\theta)$ entonces θ es la longitud de p en el sentido geográfico y ϕ es la "colatitud", es decir, el ángulo respecto al polo norte. Los meridianos y paralelos coinciden con los geográficos. La carta no cubre los polos, aunque girando la esfera obtenemos otra carta similar que los cubra. ∎

Ejemplo: Producto de variedades Si $S_1\subset\mathbb{R}^{m_1}$ y $S_2\subset\mathbb{R}^{m_2}$ son variedades entonces $S_1\times S_2\subset\mathbb{R}^{m_1+m_2}$ es también una variedad. Si $X_1:U_1\longrightarrow V_1\cap S_1$ es una carta alrededor de un punto $p_1\in S_1$ y $X_2:U_2\longrightarrow V_2\cap S_2$ es una carta alrededor de $p_2\in S_2$, entonces $X_1\times X_2:U_1\times U_2\longrightarrow(V_1\times V_2)\cap(S_1\times S_2)$ dada por $(X_1\times X_2)(u_1,u_2)=\big(X_1(u_1),X_2(u_2)\big)$ es una carta alrededor de (p_1,p_2).

Sean $\pi_i:S_1\times S_2\longrightarrow S_i$ las proyecciones. Si la carta X_1 tiene coordenadas $x_1,\ldots,x_{n_1}$ y la carta X_2 tiene coordenadas $y_1,\ldots,y_{n_2}$, entonces las coordenadas de $X_1\times X_2$ son las funciones $\pi_1\circ x_i$ y $\pi_2\circ y_i$, a las que podemos seguir llamando x_i e y_i sin riesgo de confusión. ∎

5.2 Espacios tangentes, diferenciales

Al principio de la sección anterior anticipábamos que los sistemas de coordenadas en una variedad son un análogo a los sistemas de coordenadas en un espacio afín. La diferencia principal es que en el caso afín las coordenadas están definidas sobre todo el espacio, mientras que en una variedad las tenemos definidas sólo en un entorno de cada punto. En esta sección desarrollaremos esta analogía mostrando que toda variedad diferenciable se confunde en un entorno de cada punto con una variedad afín. Para empezar, si p es un punto de una variedad S, X es una carta alrededor de p y x es su sistema de coordenadas

asociado, sabemos que en un entorno de $x(p)$ el punto $X(x)$ se confunde con $p + dX\big(x(p)\big)\big(x - x(p)\big)$, con lo que los puntos de S se confunden con los de $p + dX\big(x(p)\big)[\mathbb{R}^n]$.

Definición 5.8 Sea $S \subset \mathbb{R}^m$ una variedad diferenciable de dimensión n y sea $X : U \longrightarrow S$ una carta alrededor de un punto $p \in S$. Sea $x \in U$ tal que $X(x) = p$. Llamaremos *espacio tangente* a S en p a la variedad lineal $T_p(S) = dX(x)[\mathbb{R}^n]$. Llamaremos *variedad tangente* a S por p a la variedad afín $p+T_p(S)$.

Puesto que $JX(x)$ tiene rango n, es claro que las variedades tangentes tienen dimensión n. El teorema 5.5 prueba que el espacio tangente no depende de la carta con la que se construye, pues si X e Y son dos cartas alrededor de p, digamos $X(x) = Y(y) = p$, sabemos que $g = X \circ Y^{-1}$ es diferenciable en un entorno de x y $X = g \circ Y$, luego $dX(x) = dg(x) \circ dY(y)$, luego $dX(x)$ y $dY(y)$ tienen la misma imagen (pues $dg(x)$ es un isomorfismo). El teorema siguiente muestra más explícitamente que $T_p(S)$ sólo depende de S.

Teorema 5.9 *Sea $S \subset \mathbb{R}^m$ una variedad diferenciable de dimensión n. Entonces $T_p(S)$ está formado por el vector nulo más los vectores tangentes en p a todas las curvas regulares que pasan por p contenidas en S.*

Demostración: Sea $X : U \longrightarrow S$ una carta alrededor de p. Podemos suponer que es de la forma $X(x) = \big(x, f(x)\big)$, para una cierta función diferenciable f. Sea $X(p_1) = p$. Sea $v \in \mathbb{R}^n$ no nulo. Consideremos la curva $x(t) = p_1 + tv$. Para valores suficientemente pequeños de t se cumple que $x(t) \in U$. Consideremos la curva $\alpha(t) = X(x(t))$. Claramente α está contenida en S y cumple $\alpha(0) = p$. Su vector tangente en p es

$$\alpha'(0) = dX\big(x(0)\big)(x'(0)) = dX(p_1)(v).$$

Esto prueba que todo vector de $T_p(S)$ es de la forma indicada. Recíprocamente, si $\alpha(t)$ es una curva regular contenida en S que pasa por p, digamos $\alpha(t_0) = p$, sea $x(t) = X^{-1}(\alpha(t))$, definida en un entorno de t_0. Se cumple que $x(t)$ es derivable, pues X^{-1} no es más que la restricción de la proyección $\pi : \mathbb{R}^m \longrightarrow \mathbb{R}^n$, que es diferenciable, luego $x = \alpha \circ \pi$. Tenemos $\alpha = x \circ X$, luego $\alpha'(t) = dX\big(x(t)\big)(x'(t))$. Esta relación prueba que $x'(t) \neq 0$ o de lo contrario también se anularía $\alpha'(t)$. Por lo tanto x es regular. Además la tangente de α en p es $\alpha'(t_0) = dX(p_1)\big(x'(t_0)\big) \in T_p(S)$. ■

En la prueba de este teorema hemos visto un hecho importante: si α es una curva contenida en una variedad S y pasa por un punto p, dada una carta $X : U \longrightarrow S$ alrededor de p, podemos trasladar a la carta el arco de curva alrededor de p, es decir, existe otra curva x en U de modo que $\alpha = x \circ X$ (en un entorno de las coordenadas de p). En otras palabras, x es la representación de α en el mapa de S determinado por X.

Ejercicio: Probar que el plano tangente a una gráfica vista como variedad diferenciable coincide con el que ya teníamos definido.

Precisemos la interpretación geométrica de la variedad tangente. Ya hemos justificado que los puntos de S se confunden con los de la variedad tangente $T_p(S)$ en un entorno de p, pero más exactamente, si X es una carta alrededor de p y x es su sistema de coordenadas, hemos visto que cada punto $q \in S$ suficientemente próximo a p se confunde con el punto

$$p + dX\big(x(p)\big)\big(x(q) - x(p)\big) \in p + T_p(S).$$

Definición 5.10 Sea $S \subset \mathbb{R}^m$, $p \in S$, sea $X : U \longrightarrow V \cap S$ una carta alrededor de p y sea x su sistema de coordenadas. Llamaremos *proyección* asociada a X a la aplicación $\pi_p : S \cap V \longrightarrow T_p(S)$ dada por $\pi_p(q) = dX\big(x(p)\big)\big(x(q) - x(p)\big)$.

Según hemos visto, la interpretación geométrica de estas proyecciones consiste en que el paso $q \mapsto \pi_p(q)$ es imperceptible si tomamos puntos q suficientemente próximos a p. Ahora veamos que las coordenadas de q en la carta coinciden con las coordenadas de $\pi_p(q)$ asociadas a un cierto sistema de referencia afín en $T_p(S)$.

Sea $X : U \longrightarrow S$ una carta de una variedad S. Sea $X(x) = p$. Entonces $dX(x) : \mathbb{R}^n \longrightarrow T_p(S)$ es un isomorfismo. Por consiguiente, si $e_1, \ldots, e_n$ son los vectores de la base canónica en $\mathbb{R}^n$, sus imágenes $dX(x)(e_i) = D_iX(x)$ forman una base de $T_p(S)$. El espacio tangente no tiene una base canónica pero, según acabamos de ver, cada carta alrededor de p determina una base en $T_p(S)$. Es claro que si $q \in S$ está en el entorno de p cubierto por la carta, las coordenadas de $\pi_p(q)$ en la base asociada en $T_p(S)$ son $x(q) - x(p)$, luego si con dicha base formamos un sistema de referencia afín en $p + T_p(S)$ cuyo origen sea el punto $O = p - dX\big(x(p)\big)\big(x(p)\big)$, tenemos que las coordenadas de $\pi_p(q)$ en este sistema son precisamente $x(q)$. Cuando hablemos del sistema de referencia afín asociado a la carta nos referiremos a éste. En conclusión, cada punto q de un entorno de p en S se confunde con el punto $\pi_p(q)$ de idénticas coordenadas afines en la variedad tangente $p + T_p(S)$.

Ejercicio: Probar que si S_1 y S_2 son variedades diferenciables y $(p, q) \in S_1 \times S_2$ entonces $T_{(p,q)}(S_1 \times S_2) = T_p(S_1) \times T_q(S_2)$.

Seguidamente generalizamos la noción de diferenciabilidad al caso de aplicaciones entre variedades cualesquiera (no necesariamente abiertos de $\mathbb{R}^n$).

Definición 5.11 Diremos que una aplicación continua $f : S \longrightarrow T$ entre dos variedades es *diferenciable* (de clase C^q) en un punto $p \in S$ si existen cartas X e Y alrededor de p y $f(p)$ respectivamente de modo que $X \circ f \circ Y^{-1}$ sea diferenciable (de clase C^q) en $X^{-1}(p)$.

El teorema 5.5 implica que la diferenciabilidad de f en p no depende de la elección de las cartas X e Y, en el sentido de que si unas cartas prueban que f es diferenciable, otras cualesquiera lo prueban igualmente.

Es fácil ver que la composición de aplicaciones diferenciables es diferenciable. Una aplicación $f : U \longrightarrow \mathbb{R}^m$ definida en un abierto U de $\mathbb{R}^n$ es diferenciable

en el sentido que ya teníamos definido si y sólo si lo es considerando a U y a $\mathbb{R}^m$ como variedades diferenciables (con la identidad como carta).

Por el teorema 5.4, si $S \subset T \subset \mathbb{R}^m$ son variedades diferenciables, la inclusión $i : S \longrightarrow T$ es diferenciable, lo que se traduce en que las restricciones a S de las funciones diferenciables en T son funciones diferenciables en S.

Es claro que todas estas propiedades valen también si sustituimos la diferenciabilidad por la propiedad de ser de clase C^q.

Una aplicación $f : S \longrightarrow T$ entre dos variedades es un *difeomorfismo* si es biyectiva, diferenciable y su inversa es diferenciable. Dos variedades son *difeomorfas* si existe un difeomorfismo entre ellas.

Es obvio que las cartas de una variedad son difeomorfismos en su imagen. Más aún:

Teorema 5.12 *Todo difeomorfismo entre un abierto de $\mathbb{R}^n$ y un abierto de una variedad S en $\mathbb{R}^m$ es una carta para S.*

DEMOSTRACIÓN: Sea $f : U \longrightarrow W$ un difeomorfismo, donde $W \subset S$ es abierto en S. Entonces existe un abierto V en $\mathbb{R}^m$ tal que $f[U] = W = V \cap S$. Obviamente df tiene rango máximo en cada punto, con lo que se cumple la definición de carta. ■

En particular tenemos que las coordenadas $x_i : S \cap V \longrightarrow \mathbb{R}$ asociadas a una carta X son funciones diferenciables (son la composición de X^{-1} con las proyecciones $\pi_i : \mathbb{R}^n \longrightarrow \mathbb{R}$). Ahora definimos la diferencial de una función diferenciable.

Supongamos que $f : S \longrightarrow T$ es una aplicación entre dos variedades diferenciable en un punto p. Sean $X : U \longrightarrow S$ e $Y : W \longrightarrow T$ cartas alrededor de p y $f(p)$. Digamos que $U(x) = p$, $Y(y) = f(p)$. Entonces $j = X \circ f \circ Y^{-1}$ es diferenciable en x y tenemos las aplicaciones lineales siguientes:

$$\begin{array}{ccc} T_p(S) & & T_{f(p)}(T) \\ {\scriptstyle dX(x)}\uparrow & & \uparrow{\scriptstyle dY(y)} \\ \mathbb{R}^n & \xrightarrow{dj(x)} & \mathbb{R}^m \end{array}$$

Las flechas verticales representan isomorfismos, luego podemos definir la *diferencial* de f en p como la aplicación lineal $df(p) : T_p(S) \longrightarrow T_{f(p)}(T)$ dada por $df(p) = dX(x)^{-1} \circ dj(x) \circ dY(y)$. Teniendo en cuenta que las diferenciales aproximan localmente a las funciones correspondientes no es difícil convencerse de que $df(p)$ se confunde con f cuando los puntos de $T_p(S)$ se confunden con los de S. El teorema siguiente prueba que $df(p)$ no depende de la elección de las cartas X e Y.

Teorema 5.13 *Sea $f : S \longrightarrow T$ una aplicación diferenciable en un punto $p \in S$. Sea $v \in T_p(S)$. Si α es cualquier curva contenida en S que pase por p con tangente v, entonces $\alpha \circ f$ es una curva contenida en T que pasa por $f(p)$ con tangente $df(p)(v)$.*

DEMOSTRACIÓN: Sean X e Y cartas alrededor de p y $f(p)$ respectivamente. Digamos que $X(x) = p$ e $Y(y) = f(p)$. Sea β la representación de α en la carta X, es decir, $\alpha = \beta \circ X$. Entonces $v = \alpha'(t_0) = dX(x)\big(\beta'(t_0)\big)$.

Podemos descomponer $\alpha \circ f = \alpha \circ X^{-1} \circ X \circ f \circ Y^{-1} \circ Y$. Con la notación que hemos empleado en la definición de $df(p)$ tenemos $\alpha \circ f = \beta \circ j \circ Y$. Esto prueba que $\alpha \circ f$ es derivable en t_0 y además

$$(\alpha \circ f)'(t_0) = dY(y)\Big(dj(x)\big(\beta'(t_0)\big)\Big) = dY(y)\Big(dj(x)\big(dX(x)^{-1}(v)\big)\Big) = df(p)(v).$$

■

Es inmediato comprobar que la regla de la cadena sigue siendo válida para aplicaciones diferenciables entre variedades, es decir,

$$d(f \circ g)(p) = df(p) \circ dg\big(f(p)\big).$$

De aquí se sigue en particular que si f es un difeomorfismo, entonces $df(p)$ es un isomorfismo y $df^{-1}\big(f(p)\big) = df(p)^{-1}$.

Si $S \subset T \subset \mathbb{R}^m$ son variedades diferenciables entonces el teorema anterior prueba que la diferencial de la inclusión $i : S \longrightarrow T$ en cada punto $p \in S$ es simplemente la inclusión de $T_p(S)$ en $T_p(T)$. De aquí se sigue que la diferencial en un punto p de la restricción a S de una función f diferenciable en T es simplemente la restricción de $df(p)$ a $T_p(S)$, pues la restricción no es más que la composición con la inclusión.

Si $X : U \longrightarrow S$ es una carta de una variedad S alrededor de un punto p, entonces sus coordenadas asociadas x_i son ciertamente diferenciables. Más concretamente, si $\pi_i : \mathbb{R}^n \longrightarrow \mathbb{R}$ es la proyección en la i-ésima coordenada, tenemos que $x_i = X^{-1} \circ \pi_i$, luego $dx_i(p) = dX(p)^{-1} \circ d\pi_i(x)$ y en particular

$$dx_i(p)\,(D_j X(x)) = d\pi_i(x)(e_j) = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

es decir, las aplicaciones $dx_i(p)$ forman la base dual de $D_1X(x), \ldots, D_nX(x)$. Por consiguiente, para cada $v \in T_p(S)$ se cumple que $dx_i(p)(v)$ es la coordenada correspondiente a $D_iX(x)$ en la expresión de v como combinación lineal de las derivadas de X.

Ejemplo Consideremos el plano tangente a $\mathbb{R}^2$ en el punto $p = (1, 1)$ (que es el propio $\mathbb{R}^2$). La base asociada a la carta identidad es simplemente la base canónica (e_1, e_2), y su base dual es la dada por las proyecciones $dx(p)$, $dy(p)$. También podemos considerar también la carta determinada por las coordenadas polares (ρ, θ), es decir, $(x, y) = (\rho\cos\theta, \rho\,\mathrm{sen}\,\theta)$. Su base asociada es la formada por las derivadas parciales:

$$v_1(\rho, \theta) = (\cos\theta, \mathrm{sen}\,\theta), \quad v_2(\rho, \theta) = (-\rho\,\mathrm{sen}\,\theta, \rho\cos\theta).$$

En particular, en el punto $(1,1)$ queda

$$v_1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}\right), \quad v_2 = (-1,1).$$

Dado un vector $u \in \mathbb{R}^2$, sus coordenadas (en la base canónica) son $\big(dx(p)(u), dy(p)(u)\big)$, mientras que $\big(d\rho(p)(u), d\theta(p)(u)\big)$ son sus coordenadas en la base (v_1, v_2).

Conocemos la relación entre las diferenciales:

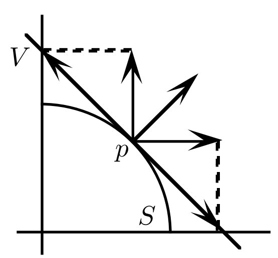
$$dx = \cos\theta\, d\rho - \rho \operatorname{sen}\theta\, d\theta, \quad dy = \operatorname{sen}\theta\, d\rho + \rho\cos\theta\, d\theta.$$

Concretamente, en el punto $(1,1)$ se cumple

$$dx = \frac{\sqrt{2}}{2}\, d\rho - d\theta, \quad dy = \frac{\sqrt{2}}{2}\, d\rho + d\theta. \tag{5.1}$$

Sea ahora S la circunferencia de radio $\sqrt{2}$. Tres posibles cartas de S alrededor de $(1,1)$ son

$$g_1(x) = \big(x, \sqrt{2-x^2}\,\big), \quad g_2(y) = \big(\sqrt{2-y^2}, y\big), \quad g_3(\theta) = \sqrt{2}\,(\cos\theta, \operatorname{sen}\theta).$$



Sus funciones coordenadas son respectivamente (las restricciones de) las funciones x, y, θ, luego sus diferenciales asociadas son las restricciones de las diferenciales correspondientes, que seguiremos llamando $dx(p)$, $dy(p)$, $d\theta(p)$. Es fácil ver que las bases asociadas a las tres cartas son respectivamente

$$v_x = (1,-1), \quad v_y = (-1,1), \quad v_\theta = (-1,1).$$

Obviamente no podemos tomar a ρ como coordenada, pues ρ es constante en S. Esto se traduce en que $d\rho(p) = 0$ (sobre $T_p(S)$). Alternativamente, vemos que los vectores de $T_p(S)$ tienen nula la primera coordenada de su expresión en la base (v_1, v_2). Como consecuencia, de (5.1) se sigue ahora que $dy = d\theta = -dx$. ■

Ejemplo Consideremos el toro T de carta

$$X(u,v) = (R\cos v + r\cos u\cos v, R\operatorname{sen} v + r\cos u \operatorname{sen} v, r\operatorname{sen} u).$$

Ya hemos comentado que X no es exactamente una carta de T, sino que las cartas de T son restricciones de X a dominios adecuados. Consideremos la circunferencia unidad $S^1 = \{x \in \mathbb{R}^2 \mid \|x\| = 1\}$. Entonces la aplicación $f : S^1 \times S^1 \longrightarrow T$ dada por

$$f(x,y) = (Ry_1 + rx_1y_1, Ry_2 + rx_1y_2, rx_2)$$

es un difeomorfismo. Notemos que si $x = (\cos u, \operatorname{sen} u)$, $y = (\cos v, \operatorname{sen} v)$, entonces $f(x, y) = X(u, v)$. Teniendo esto en cuenta es fácil ver que f es biyectiva. Además es diferenciable porque sus funciones coordenadas son polinómicas (es la restricción de una función diferenciable en $\mathbb{R}^4$). En un entorno de cada punto de T, la función f^{-1} puede expresarse como $(\cos u, \operatorname{sen} u, \cos v, \operatorname{sen} v)$, donde u, v son las funciones coordenadas de la carta de T alrededor de punto obtenida por restricción de X. Por consiguiente f es un difeomorfismo. ■

Ejercicio: Probar que un cilindro es difeomorfo al producto de un segmento por una circunferencia y que una bola abierta menos su centro es difeomorfa al producto de un segmento por una esfera.

Definición 5.14 Sea $f : S \longrightarrow \mathbb{R}$ una función definida sobre una variedad y sea $p \in S$ un punto donde f sea diferenciable. Sea X una carta de S alrededor de p y sean $x_1, \ldots, x_n$ sus coordenadas asociadas. Definimos la *derivada parcial* de f respecto a x_i en p como

$$\frac{\partial f}{\partial x_i}(p) = df(p)\big(D_i X(x)\big),$$

donde x es el vector de coordenadas de p en la carta dada.

Es claro que esta noción de derivada parcial generaliza a la que ya teníamos para el caso de funciones definidas en abiertos de $\mathbb{R}^n$. En el caso general sea $j = X \circ f$. Según la definición de $df(p)$ resulta que

$$\frac{\partial f}{\partial x_i}(p) = dj(x)(e_i) = \frac{\partial j}{\partial x_i}(x),$$

donde e_i es el i-ésimo vector de la base canónica de $\mathbb{R}^n$. Si f es diferenciable en un entorno de p tenemos

$$\frac{\partial f}{\partial x_i} = X^{-1} \circ \frac{\partial j}{\partial x_i}.$$

Ahora es claro que una función f es de clase C^q en S si y sólo si tiene derivadas parciales continuas de orden q.

Puesto que $dx_1(p), \ldots, dx_n(p)$ es la base dual de $D_1X(x), \ldots, D_nX(x)$, de la propia definición de derivada parcial se sigue que

$$df = \frac{\partial f}{\partial x_1} dx_1 + \cdots + \frac{\partial f}{\partial x_x} dx_n.$$

También es fácil ver que las reglas usuales de derivación de sumas y productos siguen siendo válidas, así como el teorema de Schwarz. Además

$$\frac{\partial x_j}{\partial x_i} = \begin{cases} 1 & \text{si } i = j \\ 0 & \text{si } i \neq j. \end{cases}$$

Es importante observar que la derivada de una función f respecto a una coordenada x_i no depende sólo de f y x_i, sino de la carta de la cual forma

parte x_i. Por ejemplo, si en la esfera de centro 0 y radio 1 consideramos un punto cuyas tres coordenadas (x, y, z) sean no nulas, en un entorno podemos considerar la carta de coordenadas (x, y), respecto a la cual

$$\frac{\partial z}{\partial x} = -\frac{x}{\sqrt{1 - x^2 - y^2}}.$$

Sin embargo, también podemos considerar la carta de coordenadas (x, z) y entonces resulta que

$$\frac{\partial z}{\partial x} = 0.$$

5.3 La métrica de una variedad

Todas las propiedades métricas de $\mathbb{R}^n$ se derivan de su producto escalar, que es una forma bilineal $\mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$. En una variedad $S \subset \mathbb{R}^m$ no tenemos definido un producto escalar, pero sí tenemos uno en cada uno de sus espacios tangentes: la restricción del producto escalar en $\mathbb{R}^m$.

Conviene introducir ciertos hechos básicos sobre formas bilineales. Puesto que son puramente algebraicas las enunciaremos para un espacio vectorial arbitrario E, pero en la práctica E será siempre el espacio tangente $T_p(S)$ de una variedad S en un punto p. Fijada una base $(v_1, \ldots, v_n)$ de E, representaremos su base dual por $(dx_1, \ldots, dx_n)$. Esta notación —puramente formal en un principio— se ajusta al único ejemplo que nos interesa, pues si $E = T_p(S)$ y $(v_1, \ldots, v_n)$ es la base asociada a una carta X, entonces la base dual que en general hemos llamado $(dx_1, \ldots, dx_n)$ es concretamente la formada por las diferenciales $dx_1(p), \ldots, dx_n(p)$, donde $x_1, \ldots, x_n$ son las funciones en S que a cada punto le asignan sus coordenadas respecto a X.

Definición 5.15 Sea E un espacio vectorial de dimensión n. Llamaremos $B(E)$ al conjunto de todas las formas bilineales $F : E \times E \longrightarrow \mathbb{R}$, que es claramente un espacio vectorial con la suma y el producto definidos puntualmente.[3]

Si $f, g : E \longrightarrow \mathbb{R}$ son aplicaciones lineales, definimos su *producto tensorial* como la forma bilineal $f \otimes g \in B(E)$ dada por $(f \otimes g)(u, v) = f(u)g(v)$.

Las propiedades siguientes son inmediatas:

a) $f \otimes (g + h) = f \otimes g + f \otimes h, \quad (f + g) \otimes h = f \otimes h + g \otimes h.$

b) $(\alpha f) \otimes g = f \otimes (\alpha g) = \alpha(f \otimes g), \quad$ para $\alpha \in \mathbb{R}$.

Teorema 5.16 *Todo elemento de $B(E)$ se expresa de forma única como*

$$F = \sum_{i,j=1}^{n} \alpha_{ij}\, dx_i \otimes dx_j, \quad \text{con } \alpha_{ij} \in \mathbb{R}.$$

Concretamente $\alpha_{ij} = F(v_i, v_j)$.

[3]Los elementos de $B(E)$ se llaman *tensores dos veces covariantes,* pero aquí no vamos a entrar en el cálculo tensorial.

DEMOSTRACIÓN: Basta observar que

$$(dx_i \otimes dx_j)(v_r, v_s) = \begin{cases} 1 & \text{si } i = r,\ j = s \\ 0 & \text{en caso contrario.} \end{cases}$$

De aquí se sigue que F y el miembro derecho de la igualdad actúan igual sobre todos los pares de vectores básicos. La unicidad es clara. ■

Por ejemplo, en estos términos el producto escalar en $\mathbb{R}^n$ viene dado por

$$dx_1 \otimes dx_1 + \cdots + dx_n \otimes dx_n.$$

Definición 5.17 Un *campo tensorial* (dos veces covariante) en una variedad $S \subset \mathbb{R}^m$ es una aplicación que a cada $p \in S$ le hace corresponder una forma bilineal en $T_p(S)$. El *tensor métrico* de S es el campo g que a cada punto p le asigna la restricción a $T_p(S)$ del producto escalar en $\mathbb{R}^m$.

Si llamamos $T(S)$ al conjunto de los campos tensoriales en S según la definición anterior, es claro que se trata de un espacio vectorial con las operaciones definidas puntualmente. Más aún, podemos definir el producto de una función $f : S \longrightarrow \mathbb{R}$ por un campo $F \in T(S)$ como el campo $fF \in T(S)$ dado por $(fF)(p) = f(p)F(p)$.

Sea $X : U \longrightarrow S$ una carta de S. Representaremos por $x_1, \ldots, x_n$ las funciones coordenadas respecto a X. Si $x \in U$ y $p = X(x)$, sabemos que $D_1X(x), \ldots, D_nX(x)$ es una base de $T_p(S)$ y $dx_1(p), \ldots, dx_n(p)$ es su base dual. Por consiguiente, todo $w \in T_p(S)$ se expresa como

$$w = dx_1(p)(w)D_1X(x(p)) + \cdots + dx_n(p)(w)D_nX(x(p)).$$

Así pues, si w_1, $w_2 \in T_p(S)$, su producto escalar es

$$g_p(w_1, w_2) = \sum_{i,j=1}^{n} D_iX(x(p))D_jX(x(p))dx_i(p)(w_1)dx_j(p)(w_2),$$

luego

$$g_p = \sum_{i,j=1}^{n} g_{ij}(p)dx_i(p) \otimes dx_j(p), \quad \text{con } g_{ij}(p) = D_iX(x(p))D_jX(x(p)),$$

o, más brevemente, como igualdad de campos:

$$g = \sum_{i,j=1}^{n} g_{ij}dx_i \otimes dx_j, \tag{5.2}$$

Esta expresión recibe el nombre de *expresión en coordenadas* del tensor métrico de S en la carta X. Las funciones g_{ij} se llaman *coeficientes* del tensor métrico en la carta dada. Claramente son funciones diferenciables. Notemos que la expresión coordenada no está definida en toda la variedad S, sino sólo sobre los puntos del rango V de la carta X.

La matriz $(g_{ij}(p))$ es la matriz del producto escalar de $T_p(S)$ en una cierta base. Es claro entonces que su determinante es no nulo. Este hecho será relevante en varias ocasiones.

A través del difeomorfismo $X : U \longrightarrow V$ juntamente con los isomorfismos $dX(x) : \mathbb{R}^n \longrightarrow T_p(S)$ podemos transportar la restricción a V del tensor métrico de S hasta un campo tensorial en U, concretamente el dado por

$$h_X(w_1, w_2) = g_p\big(dX(x)(w_1), dX(x)(w_2)\big)$$

$$= \sum_{i,j=1}^{n} g_{ij}(X(x))\, dx_i(X(x))(dX(x)(w_1))\, dx_j(X(x))(dX(x)(w_2))$$

$$= \sum_{i,j=1}^{n} g_{ij}(X(x))\, d(X \circ x_i)(x)(w_1)\, d(X \circ x_j)(x)(w_2)$$

$$= \sum_{i,j=1}^{n} g_{ij}(x)\, dx_i(x)(w_1)\, dx_j(x)(w_2),$$

donde en el último término x_i es simplemente la proyección en la i-ésima coordenada de U y $g_{ij}(x) = (X \circ g_{ij})(x)$. Por lo tanto h_X tiene la misma expresión (5.2) interpretando convenientemente las funciones.

Al transportar a la carta el tensor métrico, podemos calcular el producto de dos vectores tangentes a dos curvas α y β que se cortan en p a partir de sus representaciones en X. Digamos que $\alpha(t) = X\big(x(t)\big)$ y $\beta(t) = X\big(\bar{x}(t)\big)$ y supongamos que en t_0 pasan por p. Entonces

$$g_p(\alpha'(t_0), \beta'(t_0)) = h_X(x'(t_0), \bar{x}'(t_0)).$$

Del mismo modo que el tensor métrico de una variedad S asigna a cada punto p el producto escalar de $T_p(S)$, también podemos considerar la aplicación que a cada punto p le asigna la norma en $T_p(S)$. Ésta recibe el nombre de *elemento de longitud* de S y se representa por ds. Así pues,

$$ds(p)(v) = \|v\| = \sqrt{g_p(v, v)}.$$

El tensor métrico y el elemento de longitud se determinan mutuamente por la relación

$$ds^2(p)(u + v) = ds^2(p)(u) + ds^2(p)(v) + 2g_p(u, v),$$

luego en la práctica es equivalente trabajar con uno o con otro y ds suele dar lugar a expresiones más simples. Por ejemplo, la expresión de ds^2 en una carta es

$$ds^2 = \sum_{i,j=1}^{n} g_{ij} du_i du_j. \tag{5.3}$$

La misma expresión es válida para el campo que resulta de transportarlo al dominio de la carta interpretando adecuadamente las funciones.

El nombre de elemento de longitud se debe a que si $\alpha : [a,b] \longrightarrow S$ es una curva regular cuya imagen está contenida en el rango de una carta X y $\alpha(t) = X(x(t))$, entonces $ds^2(x'(t)) = \|\alpha'(t)\|^2$, luego la longitud de α es

$$\begin{aligned} L &= \int_a^b \|\alpha'(t)\|\, dt = \int_a^b \sqrt{\sum_{i,j=1}^n g_{ij}(x(t))x_i'(t)x_j'(t)}\, dt \\ &= \int_a^b \sqrt{\sum_{i,j=1}^n g_{ij}(x(t))\, x_i'(t)dt\, x_j'(t)dt} = \int_a^b ds, \end{aligned}$$

entendiendo ahora que en (5.3) $x = x(t)$ y $dx_i = x'(t)dt$.

En el caso de una superficie $S \subset \mathbb{R}^3$ es costumbre representar las derivadas parciales de una carta $X(u,v)$ mediante X_u, X_v y los coeficientes del tensor métrico como $E = X_uX_u$, $F = X_uX_v$, $G = X_vX_v$, de modo que la expresión en coordenadas del tensor métrico es

$$E\, du \otimes du + F(du \otimes dv + dv \otimes du) + G\, dv \otimes dv. \tag{5.4}$$

El elemento de longitud es

$$ds^2 = E\, du^2 + 2F\, dudv + G\, dv^2.$$

Ejemplo Vamos a calcular los coeficientes del tensor métrico de la superficie de revolución dada por

$$X = \big(r(u)\cos v, r(u)\operatorname{sen} v, z(u)\big).$$

Tenemos

$$\begin{aligned} X_u &= \big(r'(u)\cos v, r'(u)\operatorname{sen} v, z'(u)\big) \\ X_v &= \big(-r(u)\operatorname{sen} v, r(u)\cos v, 0\big), \end{aligned}$$

luego

$$E = r'(u)^2 + z'(u)^2, \quad F = 0, \quad G = r(u)^2.$$

Observemos que E es el módulo al cuadrado de la curva que genera la superficie, luego si su parametrización es la natural tenemos simplemente $E = 1$.

En el caso del toro tenemos $\big(r(u), z(u)\big) = (R + r\cos u, r\operatorname{sen} u)$, luego

$$E = r^2, \quad F = 0, \quad G = (R + r\cos u)^2.$$

Por lo tanto la longitud de una curva que sobre la carta venga dada por $\big(u(t), v(t)\big)$ se calcula integrando

$$ds^2 = r^2\, du^2 + (R + r\cos u)^2\, dv^2.$$

Por ejemplo, la longitud de un arco de paralelo (u_0, t), donde $t \in [0, k]$ es

$$\int_0^k ds = \int_0^k (R + r \cos u_0)\, dt = (R + r \cos u_0)k,$$

como era de esperar, dado que el paralelo es un arco de circunferencia de radio $R + r \cos u_0$. ■

Ejercicio: Calcular los coeficientes del tensor métrico del cilindro, el cono y la esfera.

Definición 5.18 Diremos que un difeomorfismo $f : S \longrightarrow T$ entre dos variedades es una *isometría* si para todo arco α contenido en S se cumple que $\alpha \circ f$ tiene la misma longitud.[4]

Explícitamente, si f es una isometría y $\alpha : [a, b] \longrightarrow S$ es un arco y $\alpha(t_0) = p$, entonces

$$\int_a^t \|\alpha'(x)\|\, dx = \int_a^t \|(\alpha \circ f)'(x)\|\, dx,$$

y derivando resulta

$$\|\alpha'(t_0)\| = \|(\alpha \circ f)'(t_0)\| = \|df(p)\big(\alpha'(t_0)\big)\|.$$

Ahora bien, todo vector no nulo de $T_p(S)$ es de la forma $\alpha'(t_0)$ para un cierto arco α, luego tenemos que $df(p) : T_p(S) \longrightarrow T_{f(p)}(T)$ es una isometría para todo punto p. Igualmente se prueba el recíproco.

Ejercicio: Probar que las isometrías de $\mathbb{R}^n$ en $\mathbb{R}^n$ en el sentido que acabamos de definir coinciden con las isometrías en el sentido del álgebra lineal.

Si f es una isometría, X es una carta alrededor de p con $X(x) = p$ y llamamos $Y = X \circ f$, es claro que Y es una carta alrededor de $f(p)$. Además tenemos que $D_iY(x) = dY(x)(e_i) = df(p)\big(dX(x)(e_i)\big) = df(p)(D_iX(x))$, de donde se sigue que los coeficientes del tensor métrico son iguales en ambas cartas, es decir,

$$g_{ij}(x) = D_iX(x)D_jX(x) = D_iY(x)D_jY(x).$$

Similarmente se concluye que si dos variedades tienen cartas con un mismo dominio y con los mismos coeficientes g_{ij} del tensor métrico entonces los fragmentos de superficie cubiertos por las cartas son superficies isométricas.

Ejemplo Consideremos la carta del cilindro dada por

$$X(u, v) = \big(r \cos \frac{v}{r}, r \operatorname{sen} \frac{v}{r}, u\big).$$

El elemento de longitud del cilindro es, en esta carta, $ds^2 = du^2 + dv^2$, que es exactamente la misma que la del plano con la identidad como carta. La aplicación X no es una isometría porque no es biyectiva, pero sí es una isometría local, en el sentido de que todo punto del plano tiene un entorno V de modo que la restricción de X es una isometría entre U y $X[U]$. Así pues, un cilindro es localmente isométrico a un plano. ■

[4]En el capítulo siguiente probaremos que toda curva de clase C^1 es rectificable. Consideraremos que esta definición se aplica a curvas y aplicaciones de clase C^1, con lo que siempre tendremos garantizado el carácter rectificable.

***Ejemplo** Observemos que las distintas expresiones que hemos obtenido para la longitud de un arco en el plano hiperbólico son de la forma (5.4) para ciertas funciones E, F, G. El caso más simple es el del semiplano de Poincaré, donde $E = G = 1/v^2$, $F = 0$. Sucede que los distintos modelos del plano hiperbólico se comportan como cartas de una superficie que no conocemos, pero de la que tenemos su elemento de longitud. Existe una teoría abstracta de variedades diferenciales que permite tratar como tales a espacios topológicos dotados de una "estructura diferenciable", definida adecuadamente, sin necesidad de que estén sumergidos en $\mathbb{R}^n$. El plano hiperbólico es una variedad en este sentido abstracto.

El plano elíptico casi puede considerarse como una superficie en $\mathbb{R}^3$: la esfera de radio 1. En realidad una esfera no es un plano elíptico, pues hemos de identificar los puntos antípodas. Sin embargo, un "fragmento" no demasiado grande de plano elíptico es isométrico a un fragmento de esfera. Por ello podemos considerar a las cartas de una esfera que no cubran más de una semiesfera como cartas del plano elíptico. Un tratamiento completamente riguroso requeriría el concepto abstracto de variedad diferenciable. ∎

5.4 Geodésicas

Imaginemos la superficie S de un planeta cuyos habitantes creen que es plano. Cuando éstos creen caminar en línea recta en realidad sus trayectorias son curvas, sin embargo su distinción entre rectas y curvas tiene un significado objetivo. Tratemos de explicitarlo. Sea $N_p(S)$ el *espacio normal* a S en p, es decir, el complemento ortogonal de $T_p(S)$. Consideremos una curva α contenida en S. Entonces $\alpha'(t) \in T_{\alpha(t)}(S)$. Podemos descomponer $\alpha''(t) = v_t(t) + v_n(t)$, donde $v_t(t) \in T_{\alpha(t)}(S)$ y $v_n(t) \in N_{\alpha(t)}(S)$. La descomposición es única. El vector v_n contiene la parte de la aceleración que mantiene a los habitantes del planeta pegados a su superficie (la gravedad) y es "invisible" para ellos, pues si el planeta fuera realmente plano la gravedad no curvaría sus trayectorias. El vector v_t contiene la variación de la velocidad que ellos detectan: determina si la trayectoria se curva a la izquierda o a la derecha. Ellos llaman rectas a las curvas que cumplen $v_t = 0$. A continuación desarrollamos estas ideas en un contexto más general:

Definición 5.19 Sea $S \subset \mathbb{R}^m$ una variedad de dimensión n, sea $\alpha : I \longrightarrow S$ una curva regular y $V : I \longrightarrow \mathbb{R}^m$ una función de clase C^1 tal que para todo $t \in I$ se cumpla $V(t) \in T_{\alpha(t)}(S)$. En estas condiciones diremos que V es un *campo de vectores* sobre α. Llamaremos *derivada covariante* de V en cada punto t a la proyección ortogonal de $V'(t)$ sobre $T_{\alpha(t)}(S)$. La representaremos por $DV(t)$.

En la situación que describíamos antes, el vector v_t es la derivada covariante del campo dado por $V(t) = \alpha'(t)$. Para ilustrar el caso general podemos pensar en un habitante del planeta S que camina rumbo norte con su brazo derecho apuntando hacia el noreste. Si interpretamos el brazo como un campo de vectores sobre su trayectoria, desde el punto de vista del caminante éste apunta

siempre en la misma dirección, pues él camina “recto”, es decir, sin desviarse ni hacia el este ni hacia el oeste, y su brazo forma un ángulo fijo con su dirección de avance. En otras palabras, considera que el campo vectorial es constante y su derivada es nula. Esto es falso, pues en realidad su trayectoria no es recta, sino una circunferencia y su brazo sí cambia de dirección (el único caso en que la dirección no variaría sería si apuntara al este o al oeste, con lo que siempre marcaría la dirección perpendicular al plano de la circunferencia en que se mueve). La que en realidad es nula es la derivada covariante del campo, que los habitantes confunden con la derivada total al desconocer la curvatura de su planeta.

En las condiciones de la definición anterior, sea $X : U \longrightarrow S$ una carta de S y expresemos la curva (localmente) como $\alpha(t) = X\big(x(t)\big)$. Entonces una base de $T_{\alpha(t)}(S)$ en cada punto es

$$D_1X(x(t)), \ldots, D_nX(x(t)),$$

luego podremos expresar

$$V(t) = a_1(t)D_1X(x(t)) + \cdots + a_n(t)D_nX(x(t)), \tag{5.5}$$

para ciertas funciones $a_i(t)$. Multiplicando la igualdad por $D_iX(x(t))$ se obtiene un sistema de ecuaciones lineales con coeficientes $g_{ij}(x(t))$. Como el determinante es no nulo, resolviendo el sistema concluimos que las funciones $a_i(t)$ son derivables. Entonces

$$V'(t) = \sum_{i=1}^n a_i'(t)D_iX(x(t)) + \sum_{i,j=1}^n a_i(t)D_{ij}X(x(t))x_j'(t). \tag{5.6}$$

El primer término es tangente a S, luego no se altera al tomar la proyección ortogonal. Para calcular la proyección del segundo conviene introducir un nuevo concepto:

Definición 5.20 Sea $X : U \longrightarrow S$ una carta de una variedad S. Llamaremos *símbolos de Christoffel* de S en la carta X a las funciones $\Gamma_{ij}^k : U \longrightarrow \mathbb{R}$ que cumplen

$$D_{ij}X = \sum_{k=1}^n \Gamma_{ij}^k D_kX + N_{ij}, \tag{5.7}$$

donde $N_{ij}(x) \in N_p(S)$ (con $p = X(x)$). Observemos que $\Gamma_{ij}^k = \Gamma_{ji}^k$.

Las proyecciones de las segundas parciales $D_{ij}X$ se obtienen eliminando la componente N_{ij}, con lo que al calcular la proyección de (5.6) llegamos a que la derivada covariante de V viene dada por

$$DV = \sum_{k=1}^n \Big(a_k' + \sum_{i,j=1}^n a_i\Gamma_{ij}^k x_j'\Big)D_kX. \tag{5.8}$$

Un hecho muy importante es que los símbolos de Christoffel, y por consiguiente la derivada covariante, dependen únicamente de los coeficientes g_{ij} de la primera forma fundamental de S. En efecto, multiplicando las ecuaciones (5.7) por D_lX obtenemos

$$D_{ij}XD_lX = \sum_{k=1}^{n} g_{kl}\Gamma_{ij}^k.$$

Una simple comprobación nos da que

$$D_{ij}XD_lX = \frac{1}{2}(D_ig_{jl} + D_jg_{il} - D_lg_{ij}),$$

luego en total resulta

$$\sum_{k=1}^{n} g_{kl}\Gamma_{ij}^k = \frac{1}{2}(D_ig_{jl} + D_jg_{il} - D_lg_{ij}). \tag{5.9}$$

Fijando i, j y variando l obtenemos un sistema de n ecuaciones lineales con n incógnitas y coeficientes (g_{kl}), que nos permite despejar los símbolos Γ_{ij}^k en términos de los coeficientes g_{ij} y sus derivadas, como queríamos probar. Ahora nos ocupamos con detalle del caso particular que describíamos al principio de la sección:

Definición 5.21 Sea $\alpha(t)$ una curva contenida en una variedad S. Llamaremos *aceleración geodésica*[5] de α a la derivada covariante del campo vectorial α'.

Supongamos que α está parametrizada por el arco. Entonces $\|\alpha'(s)\| = 1$, luego derivando resulta $\alpha''(s)\alpha'(s) = 0$, y esta ortogonalidad se conserva al proyectar sobre $T_p(S)$, de modo que $D\alpha'(s)$ es perpendicular al vector tangente de α. Llamaremos *curvatura geodésica* de α a $\kappa_g = \|D\alpha'\|$. Si $\kappa_g \neq 0$ definimos el *vector normal geodésico* de α como el vector $\kappa_g^{-1}D\alpha'$, de modo que $D\alpha' = \kappa_g n_g$.

En el caso de que α no esté parametrizada por el arco el vector normal geodésico y la curvatura geodésica se definen a través de su parametrización natural. Explícitamente, si $\alpha(t)$ es una curva contenida en S y $s(t)$ es su longitud de arco, usando la notación $v = s'(t) = \|\alpha'(t)\|$, $a = v'(t)$ para la velocidad y aceleración sobre la trayectoria y $T = \alpha'(s)$ para el vector tangente, tenemos

$$\alpha'(t) = v\alpha'(s), \quad \alpha''(t) = aT + v^2\alpha''(s).$$

Al proyectar sobre el espacio tangente resulta

$$D\alpha'(t) = aT + v^2\kappa_g n_g.$$

De este modo, la aceleración geodésica de α se descompone en una aceleración tangencial, cuyo módulo a es la tasa de variación de la velocidad v, y una

[5]La *geodesia* (gr. = división de la tierra) estudia la forma de la Tierra, deducida a partir de mediciones realizadas desde su superficie. La geometría diferencial ha adoptado este adjetivo para referirse en general a los conceptos que puede medir un "habitante" de una variedad arbitraria sin salir de ella.

aceleración normal, cuyo módulo es $v^2\kappa_g$. Un habitante del planeta S que "crea" vivir en $T_p(S)$ confundirá la aceleración geodésica, el vector normal geodésico y la curvatura geodésica de α con la aceleración, el vector normal y la curvatura de α. Por lo tanto llamará rectas a las curvas sin aceleración geodésica:

Definición 5.22 Una curva α contenida en una variedad S es una *geodésica*[6] si cumple $\kappa_g = 0$, o equivalentemente, si $D\alpha'$ es proporcional a α' en cada punto. En tal caso el factor de proporcionalidad es simplemente $a = s''(t)$, donde s es la longitud de arco, por lo que si α está parametrizada por el arco entonces α es una geodésica si y sólo si $D\alpha' = 0$.

Vamos a particularizar las ecuaciones que determinan la derivada covariante de un campo al caso de la aceleración geodésica de una curva. Si $\alpha(t) = X(x(t))$, entonces

$$\alpha'(t) = \sum_{i=1}^{n} D_i X(x(t)) x_i'(t),$$

luego si en (5.5) hacemos $V = \alpha'$ tenemos $a_i = x_i'$, luego la fórmula (5.8) se convierte en

$$DV = \sum_{k=1}^{n} \Big(x_k'' + \sum_{i,j=1}^{n} \Gamma_{ij}^k x_i' x_j' \Big) D_k X.$$

El vector

$$\Big(x_k'' + \sum_{i,j=1}^{n} \Gamma_{ij}^k x_i' x_j' \Big)_{k=1}^{n}$$

es la antiimagen por dX de $D\alpha'$, es decir, la representación en el mapa de la aceleración geodésica de α. Lo llamaremos *expresión en coordenadas* de dicha aceleración geodésica.

La condición necesaria y suficiente para que una curva parametrizada por el arco de coordenadas $x(s)$ sea una geodésica es

$$x_k'' + \sum_{i,j=1}^{n} \Gamma_{ij}^k x_i' x_j' = 0, \quad k = 1, \ldots, n. \tag{5.10}$$

Si la parametrización es arbitraria sólo hemos de exigir que el vector formado por los miembros izquierdos sea proporcional a x'.

Ejemplo Si una carta $X(u, v)$ de una superficie $S \subset \mathbb{R}^3$ cumple $F = 0$, las ecuaciones (5.9) se reducen a

$$\begin{aligned} \Gamma_{11}^1 &= \frac{E_u}{2E}, \quad \Gamma_{11}^2 = -\frac{E_v}{2G}, \quad \Gamma_{12}^1 = \frac{E_v}{2E}, \\ \Gamma_{12}^2 &= \frac{G_u}{2G}, \quad \Gamma_{22}^1 = -\frac{G_u}{2E}, \quad \Gamma_{22}^2 = \frac{G_v}{2G}. \end{aligned} \tag{5.11}$$

■

[6]Deberíamos decir "recta geodésica", es decir, el equivalente en S a una recta, pero es preferible contraer el término pues, al fin y al cabo, normalmente las geodésicas no son rectas.

Ejemplo En un plano (tomando como carta la identidad) todos los símbolos de Christoffel son nulos, por lo que las geodésicas parametrizadas por el arco son las curvas que cumplen $(u'', v'') = (0, 0)$, es decir, las rectas. ■

Ejemplo En la superficie de revolución generada por la curva $\big(r(u), z(u)\big)$, suponiendo a ésta parametrizada por el arco, los únicos símbolos de Christoffel no nulos son

$$\Gamma_{12}^2 = \frac{r'(u)}{r(u)}, \quad \Gamma_{22}^1 = -r(u)r'(u).$$

Por lo tanto las ecuaciones de las geodésicas parametrizadas por el arco son

$$u'' = v'^2 r(u)r'(u), \quad v'' = -2u'v'\,\frac{r'(u)}{r(u)}.$$

Es inmediato comprobar que los meridianos (t, v_0) cumplen estas ecuaciones, luego son geodésicas. Si se cumple $r'(u) = 0$, (por ejemplo en los extremos locales de r) entonces el paralelo (u_0, t) también cumple las ecuaciones, luego es una geodésica.

En el caso concreto de la esfera los meridianos son los arcos de circunferencia de radio máximo que unen los polos. Dada la simetría de la esfera, que permite tomar cualquier par de puntos antípodas como polos, podemos afirmar que todas las circunferencias máximas son geodésicas. Para una carta dada, el único paralelo (u_0, t) que cumple $r'(u_0) = 0$ es el ecuador de la esfera, que también es una circunferencia máxima, luego ya sabíamos que es una geodésica. ■

***Ejemplo** Las fórmulas que determinan los símbolos de Christoffel a partir de los coeficientes del tensor métrico hacen que tenga sentido calcularlos en el caso de los planos elíptico e hiperbólico, donde la definición de derivada covariante que hemos dado no es aplicable. El hecho de que las circunferencias máximas de una esfera sean geodésicas se traduce en que las rectas elípticas sean geodésicas del plano elíptico (pues éste es localmente isométrico a una esfera de radio 1). Veamos ahora que las rectas hiperbólicas son geodésicas del plano hiperbólico. Para ello trabajaremos con el semiplano de Poincaré, donde los símbolos de Christoffel son más sencillos. Teniendo en cuenta que $E = G = 1/v^2$ y $F = 0$ es fácil ver que los únicos símbolos no nulos son

$$\Gamma_{11}^2 = \frac{1}{v}, \quad \Gamma_{12}^1 = -\frac{1}{v}, \quad \Gamma_{22}^2 = -\frac{1}{v}.$$

La aceleración covariante de una curva de coordenadas (u, v) tiene coordenadas

$$\left(u'' - 2\,\frac{u'v'}{v},\ v'' + \frac{u'^2 - v'^2}{v}\right).$$

Para las rectas verticales $(u, v) = (u_0, t)$ la aceleración es $(0, -1/t)$, que efectivamente es proporcional a $(u', v') = (0, 1)$, luego son geodésicas. Las rectas

proyectivas restantes son las semicircunferencias $(u,v) = (u_0 + r\cos t, r \operatorname{sen} t)$. Un simple cálculo nos da que la aceleración en este caso es

$$\left(r\cos t, -r\,\frac{\cos^2 t}{\operatorname{sen} t}\right) = -\frac{\cos t}{\operatorname{sen} t}\,(-r\operatorname{sen} t, r\cos t),$$

proporcional a (u', v'), luego todas las rectas proyectivas son geodésicas. ■

5.5 Superficies

Terminaremos el capítulo con algunos resultados específicos sobre superficies $S \subset \mathbb{R}^3$. Si X es una carta de una superficie S, entonces X_u, X_v son en cada punto (u,v) una base del plano tangente en $X(u,v)$, luego el vector $X_u \wedge X_v$ es no nulo y perpendicular a dicho plano. Si llamamos α al ángulo formado por X_u y X_v en un punto dado, entonces

$$\|X_u \wedge X_v\|^2 = \|X_u\|^2\|X_v\|^2(1-\cos^2\alpha) = X_uX_u\,X_vX_v - (X_uX_v)^2 = EG - F^2.$$

Así pues, $\|X_u \wedge X_v\| = \sqrt{EG-F^2}$.

Definición 5.23 La *aplicación de Gauss* asociada a una carta $X : U \longrightarrow S$ de una superficie S es la aplicación $n : U \longrightarrow \mathbb{R}^3$ dada por

$$n(u,v) = \frac{X_u \wedge X_v}{\|X_u \wedge X_v\|} = \frac{X_u \wedge X_v}{\sqrt{EG-F^2}}.$$

De este modo, $n(u,v)$ es en cada punto un vector unitario perpendicular a S en $X(u,v)$. Esto lo determina completamente salvo en su sentido. Si cambiamos de carta, el sentido de n puede cambiar.

Si X y $\bar{X}$ son dos cartas que cubren una misma región conexa de una variedad, entonces $n(u,v) = \epsilon(u,v)\bar{n}(u,v)$, donde $\epsilon(u,v) = \pm 1$. Es claro que ϵ es una función continua en un conexo, luego ha de ser constante. En definitiva, $n(u,v) = \pm\bar{n}(u,v)$. Resulta, pues, que en un entorno de cada punto de S existen exactamente dos determinaciones opuestas del vector normal. A cualquiera de ellas la llamaremos también *aplicación de Gauss* de la superficie.

Si llamamos $G \subset S$ a la imagen de X, la aplicación n induce otra aplicación $n : G \longrightarrow S^2$, donde S^2 es la esfera de centro $(0,0,0)$ y radio 1, que está unívocamente determinada en un entorno de cada punto excepto por su signo. Es importante notar que no siempre es posible extender esta aplicación n a toda la superficie S (sin perder la continuidad).

De momento no vamos a entrar en detalles, pero la figura muestra un ejemplo de variedad sobre la cual no es posible definir un vector normal. Se la conoce como *banda de Möbius*. Es una cinta pegada por sus extremos tras haberla girado media vuelta. Si la aplicación de Gauss pudiera definirse sobre toda la banda M, al componerla con una curva $\alpha : \mathbb{R} \longrightarrow M$ que dé una

vuelta completa obtendríamos un vector normal sobre α que variaría de forma continua, pero es claro que al dar una vuelta completa el vector normal termina en sentido inverso a como empezó, cuando por continuidad debería tender al vector de partida.

La aplicación de Gauss aporta información importante sobre las superficies y simplifica algunos de los conceptos que hemos estudiado para variedades arbitrarias. Por ejemplo, en la sección anterior hemos estudiado la componente tangencial (o geodésica) de la curvatura de una curva contenida en una variedad. Del mismo modo podemos definir la curvatura normal como el módulo de la componente normal de la segunda derivada. En el caso de las superficies en $\mathbb{R}^3$ podemos apoyarnos en la aplicación de Gauss.

Definición 5.24 Sea S una superficie (al menos de clase C^2) y α una curva contenida en S parametrizada por el arco y que pase por un punto p. Fijada una determinación n del vector normal a S alrededor de p, llamaremos *curvatura normal* de α a $\kappa_n = \alpha'' n$. Definimos $N_n = \kappa_n n$ y $N_t = \alpha'' - N_n$.

Notemos que el signo de κ_n depende de la determinación que elijamos de la aplicación de Gauss. Supongamos que sobre una carta la curva es $\big(u(t), v(t)\big)$. Entonces

$$\alpha' = X_u u' + X_v v', \quad \alpha'' = X_{uu} u'^2 + X_u u'' + X_{uv} u'v' + X_{uv} u'v' + X_{vv} v'^2 + X_v v'',$$

luego

$$\kappa_n = \alpha'' n = (X_{uu} n) u'^2 + 2(X_{uv} n) u'v' + (X_{vv} n) v'^2.$$

Llamamos

$$e = X_{uu} n, \quad f = X_{uv} n, \quad g = X_{vv} n,$$

que son funciones de la carta X (salvo por el signo, que depende de la elección del sentido de n). Si la parametrización de la curva no es la natural y $s(t)$ es la longitud de arco, usamos la regla de la cadena:

$$\frac{du}{dt} = \frac{du}{ds}\frac{ds}{dt}, \quad \frac{dv}{dt} = \frac{dv}{ds}\frac{ds}{dt},$$

con la que la fórmula, llamando ahora u', v', s' a las derivadas respecto de t (hasta ahora eran las derivadas respecto de s), se convierte en

$$\kappa_n = \frac{e\, u'^2 + 2f\, u'v' + g\, v'^2}{s'^2}.$$

Observemos que esta expresión no depende de la curva (u, v), sino sólo de su derivada (u', v') (recordemos que $s' = \|(u', v')\|$). De aquí deducimos:

Teorema 5.25 (Teorema de Meusnier) *Si S es una superficie, todas las curvas contenidas en S que pasan por un punto p con un mismo vector tangente tienen la misma curvatura normal. Ésta viene dada por*

$$\kappa_n = \frac{e\, du^2 + 2f\, dudv + g\, dv^2}{E\, du^2 + 2F\, dudv + G\, dv^2}.$$

DEMOSTRACIÓN: Es claro que X es una carta alrededor de p y α es una curva contenida en S que pasa por p con vector tangente w, entonces la representación de α en la carta X es $X^{-1} \circ \alpha$, luego el vector tangente de esta representación —el que en la discusión previa al teorema llamábamos (u', v')— es $\big(du(p)(w), dv(p)(w)\big)$, donde ahora u y v son las funciones coordenadas de X. Así pues, la fórmula que habíamos obtenido nos da que

$$\kappa_n(p)(w) = \frac{e(p)\, du(p)^2(w) + 2f(p)\, du(p)(w)dv(p)(w) + g(p)\, dv(p)^2(w)}{E(p)\, du(p)^2(w) + 2F(p)\, du(p)(w)dv(p)(w) + G\, dv(p)^2(w)},$$

entendiendo aquí a e, f, g como las composiciones con X^{-1} de las funciones del mismo nombre que teníamos definidas. ■

Definición 5.26 El elemento de longitud de una superficie S se conoce también con el nombre que le dio Gauss: *la primera forma fundamental* de S. Definimos la *segunda forma fundamental* de S como la aplicación que a $p \in S$ y cada vector $w \in T_p(S)$ le asigna la curvatura normal en p de las curvas contenidas en S que pasan por p con tangente w multiplicada por $\|w\|^2$.

El teorema anterior prueba que la segunda forma fundamental es en cada punto $p$una forma cuadrática definida sobre $T_p(S)$. Concretamente, si fijamos una carta tenemos

$$F^1 = E\, du^2 + 2F\, dudv + G\, dv^2, \quad F^2 = e\, du^2 + 2f\, dudv + g\, dv^2.$$

Ambas formas cuadráticas pueden considerarse definidas tanto sobre la superficie S como sobre el dominio de la carta (en cuyo caso du y dv representan simplemente las proyecciones de $\mathbb{R}^2$). Sin embargo, una diferencia importante es que, aunque las expresiones anteriores son válidas únicamente sobre el rango de una carta, la primera forma fundamental está definida sobre toda la superficie y está completamente determinada por la misma, mientras que la segunda sólo la tenemos definida en un entorno de cada punto y además salvo signo.

Para calcular explícitamente la segunda forma fundamental de una superficie notamos que

$$e = X_{uu}n = X_{uu}\frac{X_u \wedge X_v}{\|X_u \wedge X_v\|} = \frac{(X_{uu}, X_u, X_v)}{\sqrt{EG - F^2}},$$

e igualmente

$$f = \frac{(X_{uv}, X_u, X_v)}{\sqrt{EG - F^2}}, \quad g = \frac{(X_{vv}, X_u, X_v)}{\sqrt{EG - F^2}}.$$

Ejemplo Los coeficientes de la segunda forma fundamental de la superficie de revolución generada por la curva $\big(r(u), z(u)\big)$ son

$$e = \frac{z''(u)r'(u) - z'(u)r''(u)}{\sqrt{r'(u)^2 + z'(u)^2}}, \quad f = 0, \quad g = \frac{z'(u)r(u)}{\sqrt{r'(u)^2 + z'(u)^2}}.$$

Para el caso del toro tenemos $\big(r(u), z(u)\big) = (R + r\cos v, r \operatorname{sen} v)$ luego queda

$$e = r, \quad f = 0, \quad g = R\cos u + r\cos^2 u.$$

■

Ejercicio: Comprobar que la curvatura normal en todo punto de la esfera de radio r y en toda dirección es igual a $\pm 1/r$, donde el signo es positivo si elegimos el vector normal que apunta hacia dentro de la esfera y negativo en caso contrario.

5.6 La curvatura de Gauss

Es un hecho conocido que si F es una forma bilineal simétrica en un espacio euclídeo existe una base ortonormal en la que la matriz de F es diagonal. Podemos aplicar esto a un plano tangente $T_p(S)$ de una superficie tomando el producto escalar determinado por la primera forma fundamental y como F la segunda forma fundamental. Entonces concluimos que existe una base (e_1, e_2) de $T_p(S)$ en la cual las expresión en coordenadas de las formas fundamentales es

$$F^1(x, y) = x^2 + y^2 \quad \text{y} \quad F^2(x, y) = \lambda_1 x^2 + \lambda_2 y^2.$$

Los números λ_1 y λ_2 son los valores propios de cualquiera de las matrices de F^2 en cualquier base ortonormal de $T_p(S)$, luego están unívocamente determinados salvo por el hecho de que un cambio de carta puede cambiar sus signos. Podemos suponer $\lambda_1 \leq \lambda_2$. Entonces se llaman respectivamente *curvatura mínima* y *curvatura máxima* de S en p. En efecto, se trata del menor y el mayor valor que toma F^2 entre los vectores de norma 1, pues

$$\lambda_1 = \lambda_1(x^2 + y^2) \leq \lambda_1 x^2 + \lambda_2 y^2 = F^2(x, y) \leq \lambda_2(x^2 + y^2) = \lambda_2.$$

Si $w \in T_p(S)$ tiene norma arbitraria entonces aplicamos esto a $w/\|w\|$ y concluimos que

$$\lambda_1 \leq \frac{F^2(w)}{F^1(w)} \leq \lambda_2,$$

es decir, $\lambda_1 \leq \kappa_n \leq \lambda_2$. Así pues, λ_1 y λ_2 son la menor y la mayor curvatura normal que alcanzan las curvas que pasan por p. Además se alcanzan en direcciones perpendiculares e_1 y e_2, llamadas *direcciones principales* en p. Notemos que puede ocurrir $\lambda_1 = \lambda_2$, en cuyo caso la curvatura normal es la misma en todas direcciones y no hay direcciones principales distinguidas. Los puntos de S donde $\lambda_1 = \lambda_2$ se llaman *puntos umbilicales.*

Veamos ahora cómo calcular las direcciones principales en una carta. Consideremos la fórmula de Meusnier como función (diferenciable) de dos variables. Si (du, dv) marca una dirección principal[7] entonces κ_n es máximo o mínimo

[7] Aquí podemos considerar $(du, dv) \in \mathbb{R}^2$. La notación diferencial está motivada por lo siguiente: Fijada una carta con coordenadas u, v, una curva regular en la superficie viene determinada por una representación coordenada $(u(t), v(t))$. El vector tangente a la curva en un punto dado marcará una dirección principal si y sólo si la fórmula de Meusnier evaluada en $(u'(t), v'(t))$ toma un valor máximo o mínimo, pero dicha fórmula depende sólo de las diferenciales $(du(t), dv(t))$, por lo que en realidad buscamos una relación entre du y dv.

en este punto, luego el teorema 4.15 afirma que sus derivadas parciales han de anularse en él. Así pues, se ha de cumplir

$$\begin{aligned}
\frac{\partial \kappa_n}{\partial du} &= \frac{2(e\,du + f\,dv)}{F^1(du,dv)} - \frac{2(E\,du + F\,dv)}{F^1(du,dv)^2}\,F^2(du,dv) = 0,\\
\frac{\partial \kappa_n}{\partial dv} &= \frac{2(f\,du + g\,dv)}{F^1(du,dv)} - \frac{2(F\,du + G\,dv)}{F^1(du,dv)^2}\,F^2(du,dv) = 0.
\end{aligned}$$

Despejando obtenemos

$$\begin{aligned}
\kappa_n &= \frac{F^2(du,dv)}{F^1(du,dv)} = \frac{e\,du + f\,dv}{E\,du + F\,dv}, \qquad (5.12)\\
\kappa_n &= \frac{F^2(du,dv)}{F^1(du,dv)} = \frac{f\,du + g\,dv}{F\,du + G\,dv}.
\end{aligned}$$

Al igualar ambas ecuaciones obtenemos una condición necesaria para que un vector indique una dirección principal. Es fácil ver que puede expresarse en la forma:

$$\begin{vmatrix} dv^2 & -dudv & du^2\\ E & F & G\\ e & f & g \end{vmatrix} = 0.$$

Si (E,F,G) (en un punto) es múltiplo de (e,f,g) entonces la ecuación se cumple trivialmente, pero por otra parte es claro que la curvatura normal es constante y no hay direcciones principales. En caso contrario es claro tenemos una forma cuadrática con al menos dos coeficientes no nulos. Si suponemos, por ejemplo, que el coeficiente de dv^2 es no nulo, entonces $du \neq 0$, y al dividir entre du^2 la forma cuadrática se convierte en una ecuación de segundo grado en la razón dv/du. Esta ecuación tiene a lo sumo dos soluciones linealmente independientes, luego éstas han de ser necesariamente las direcciones principales. Por consiguiente la ecuación caracteriza dichas direcciones.

Definición 5.27 Se llama *curvatura media* y *curvatura total o de Gauss* de una superficie S en un punto p a los números

$$H = \frac{\lambda_1 + \lambda_2}{2}, \quad K = \lambda_1\lambda_2.$$

Notemos que el signo de H depende de la carta, mientras que el de K es invariante. Si operamos en (5.12) obtenemos

$$\begin{aligned}
(e - E\kappa_n)du + (f - F\kappa_n)dv &= 0,\\
(f - F\kappa_n)du + (g - G\kappa_n)dv &= 0.
\end{aligned}$$

Puesto que el sistema tiene una solución no trivial en (du,dv) se ha de cumplir

$$\begin{vmatrix} e - E\kappa_n & f - F\kappa_n\\ f - F\kappa_n & g - G\kappa_n \end{vmatrix} = 0,$$

o equivalentemente

$$(EG - F^2)\kappa_n^2 - (eG - 2Ff + gE)\kappa_n + (eg - f^2) = 0.$$

Esta ecuación la cumplen las curvaturas principales $\kappa_n = \lambda_1, \lambda_2$ y por otro lado tiene sólo dos soluciones, luego

$$\begin{aligned} H &= \frac{eG - 2Ff + gE}{2(EG - F^2)}, \\ K &= \frac{eg - f^2}{EG - F^2}. \end{aligned} \tag{5.13}$$

En particular vemos que la curvatura de Gauss es el cociente de los determinantes de las dos formas fundamentales.

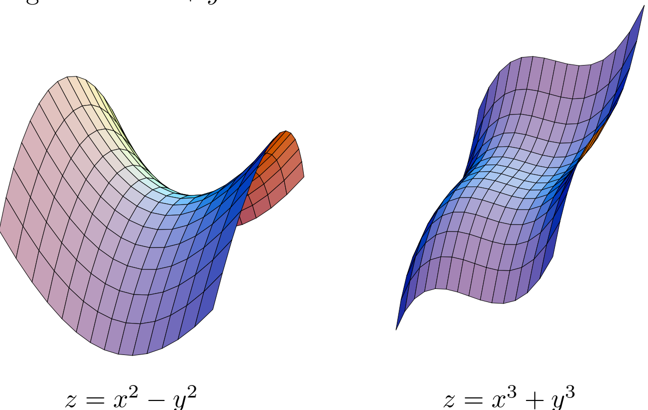
Ejercicio: Calcular la curvatura media y la curvatura de Gauss del cilindro, el cono, el toro y la esfera.

Definición 5.28 Un punto p de una superficie S es *elíptico* o *hiperbólico* según si $K(p) > 0$ o $K(p) < 0$. Si $K(p) = 0$ distinguiremos entre puntos *parabólicos,* cuando sólo una de las curvaturas extremas es nula y puntos *planos,* cuando las dos curvaturas extremas son nulas.

Si un punto es elíptico todas las curvas que pasan por él tienen la curvatura normal del mismo signo, por lo que la superficie se curva toda hacia el mismo lado del plano tangente, como es el caso de la esfera o del toro. Si un punto es hiperbólico entonces hay curvas (perpendiculares, de hecho) que pasan por él con curvaturas en sentidos opuestos, luego la superficie tiene puntos próximos a ambos lados del plano tangente. Es el caso del hiperboloide $z = x^2 - y^2$, cuya curvatura en la carta $(u, v, u^2 - v^2)$ viene dada por $K = -4/\sqrt{4u^2 + 4v^2 + 1}^{\,3}$.

Los puntos de un cilindro son parabólicos. Las curvas u =cte. y v =cte. son circunferencias de radio r y rectas, respectivamente. Las primeras tienen curvatura normal $\lambda_2 = 1/r$ y las segundas $\lambda_1 = 0$. Es fácil ver que se trata de las curvaturas principales. Todos los cálculos son sencillos.

Todos los puntos de un plano son puntos planos. Otro ejemplo es el punto $(0, 0)$ en la gráfica de $x^3 + y^3$.



$z = x^2 - y^2$ $\qquad$ $z = x^3 + y^3$

Probamos ahora una caracterización algebraica de la curvatura de Gauss que más adelante nos dará una interpretación geométrica de la misma. Observemos que si n es una determinación del vector normal alrededor de un punto p en una superficie S y llamamos S^2 a la esfera de centro $(0,0,0)$ y radio 1, entonces $dn(p) : T_p(S) \longrightarrow T_{n(p)}(S^2)$, pero como $n(p)$ es perpendicular a $T_p(S)$, en realidad $T_{n(p)}(S^2) = T_p(S)$, luego podemos considerar a $dn(p)$ como un endomorfismo de $T_p(S)$.

Teorema 5.29 *Sea S una superficie y n una determinación del vector normal alrededor de un punto p. Entonces $K(p) = |dn(p)|$.*

Demostración: Sea X una carta alrededor de p. Entonces una base de $T_p(S)$ la forman los vectores X_u y X_v. Llamemos $n(u,v)$ a $X \circ n$. Entonces

$$\begin{aligned} dn(p)(X_u) &= dn(p)(dX(u,v)(1,0)) = dn(u,v)(1,0) = n_u, \\ dn(p)(X_v) &= dn(p)(dX(u,v)(0,1)) = dn(u,v)(0,1) = n_v. \end{aligned}$$

Si expresamos

$$\begin{aligned} n_u &= aX_u + bX_v \\ n_v &= cX_u + dX_v \end{aligned}$$

entonces el determinante de $dn(p)$ es el de la matriz formada por a, b, c, d. Notemos que derivando las igualdades $nX_u = nX_v = 0$ se deduce la relación $n_uX_u = -nX_{uu} = -e$ y similarmente $n_uX_v = n_vX_u = -f$, $n_vX_v = -g$. Por consiguiente al multiplicar las ecuaciones anteriores por X_u y X_v obtenemos

$$-e = aE + bF, \quad -f = aF + bG, \quad -f = cE + dF, \quad -g = cF + dG,$$

de donde

$$-\begin{pmatrix} e & f \\ f & g \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} E & F \\ F & G \end{pmatrix}.$$

Tomando determinantes concluimos que

$$eg - f^2 = |dn(p)|\,(EG - F^2),$$

luego efectivamente $|dn(p)| = K(p)$. ■

Consideremos ahora las fórmulas (5.7) que definen los símbolos de Christoffel. Al particularizarlas al caso de una superficie se convierten en

$$\begin{aligned} X_{uu} &= \Gamma_{11}^1 X_u + \Gamma_{11}^2 X_v + en, \\ X_{uv} &= \Gamma_{12}^1 X_u + \Gamma_{12}^2 X_v + fn, \\ X_{vv} &= \Gamma_{22}^1 X_u + \Gamma_{22}^2 X_v + gn, \end{aligned}$$

(en principio la componente normal ha de ser de la forma αn para cierto α, y multiplicando la igualdad por n se sigue que $\alpha = e, f, g$ según el caso.)

De estas ecuaciones se sigue

$$\begin{aligned} X_{uu}X_{vv} - X_{uv}^2 = eg - f^2 \quad &+ \quad \big(\Gamma_{11}^1\Gamma_{22}^1 - (\Gamma_{12}^1)^2\big)E \\ &+ \quad \big(\Gamma_{11}^1\Gamma_{22}^2 + \Gamma_{11}^2\Gamma_{22}^1 - 2\Gamma_{12}^1\Gamma_{12}^2\big)F \\ &+ \quad \big(\Gamma_{11}^2\Gamma_{22}^2 - (\Gamma_{12}^2)^2\big)G. \end{aligned}$$

Por otra parte, derivando respecto a v y u respectivamente las relaciones

$$X_{uu}X_v = F_u - \frac{1}{2}E_v, \quad X_{uv}X_v = \frac{1}{2}G_u$$

y restando los resultados obtenemos

$$X_{uu}X_{vv} - X_{uv}^2 = -\frac{1}{2}E_{vv} + F_{uv} - \frac{1}{2}G_{uu}.$$

En definitiva resulta la expresión

$$\begin{aligned} eg - f^2 = -\frac{1}{2}E_{vv} + F_{uv} - \frac{1}{2}G_{uu} \quad &- \quad \big(\Gamma_{11}^1\Gamma_{22}^1 - (\Gamma_{12}^1)^2\big)E \\ &- \quad \big(\Gamma_{11}^1\Gamma_{22}^2 + \Gamma_{11}^2\Gamma_{22}^1 - 2\Gamma_{12}^1\Gamma_{12}^2\big)F \\ &- \quad \big(\Gamma_{11}^2\Gamma_{22}^2 - (\Gamma_{12}^2)^2\big)G. \end{aligned}$$

La fórmula (5.13) muestra ahora que la curvatura de Gauss de un punto depende únicamente de los coeficientes E, F, G de la primera forma fundamental y sus derivadas. Puesto que dos superficies localmente isométricas tienen cartas con los mismos coeficientes E, F, G, hemos probado el resultado que Gauss, en sus *Diquisitiones generales circa superficies curuas*, presentó con el nombre de *theorema egregium:*

Teorema 5.30 (Gauss) *Las isometrías locales conservan la curvatura.*

Las ecuaciones (5.11) nos dan la siguiente expresión para la curvatura respecto a una carta con $F = 0$:

$$K = \frac{E_uG_u + E_v^2}{4E^2G} + \frac{E_vG_v + G_u^2}{4EG^2} - \frac{E_{vv} + G_{uu}}{2EG}, \quad \text{si } F = 0.$$

Desde aquí es fácil deducir a su vez los siguientes casos particulares:

$$K = -\frac{1}{2A}\left(\frac{\partial^2 \log A}{\partial u^2} + \frac{\partial^2 \log A}{\partial v^2}\right), \quad \text{si } F = 0,\ E = G = A,$$

$$K = -\frac{1}{\sqrt{G}}\frac{\partial^2\sqrt{G}}{\partial u^2}, \quad \text{si } F = 0,\ E = 1.$$

En particular, la curvatura de la superficie de revolución definida por la curva $\big(r(u), z(u)\big)$ es

$$K = -\frac{r''(u)}{r(u)}.$$

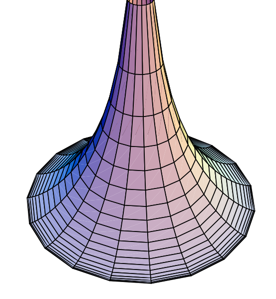
***Ejemplo** Notemos que sin el teorema de Gauss no tendría sentido hablar de la curvatura del plano hiperbólico, pues lo único que sabemos de él es que sus modelos se comportan como cartas de una variedad desconocida de la que tenemos su primera forma fundamental. Sin embargo, las fórmulas anteriores nos permiten calcular su curvatura a partir de estos datos. Por ejemplo, en el caso del semiplano de Poincaré, donde $E = G = 1/v^2$ y $F = 0$, ahora es fácil calcular que $K = -1$. Así pues, si pudiéramos identificar el plano hiperbólico con una superficie en $\mathbb{R}^3$, ésta tendría que tener curvatura constante igual a -1. En el caso del plano elíptico sabemos que localmente es como la esfera de radio 1, luego si pudiéramos identificar el plano elíptico con una superficie de $\mathbb{R}^3$, ésta tendría que tener curvatura constante igual a 1. Ahora vamos a probar que existen variedades cuya relación con el plano hiperbólico es la misma que hay entre la esfera y el plano elíptico. ■

Ejemplo Se llama *pseudoesfera* a la superficie de revolución P generada por la tractriz. Recordemos que la tractriz es

$$\big(r(u), z(u)\big) = \Big(l \operatorname{sen} u, l \cos u + l \log \tan \frac{u}{2}\Big).$$

Por lo tanto la pseudoesfera está dada por

$$X(u, v) = \Big(l \operatorname{sen} u \cos v, l \operatorname{sen} u \operatorname{sen} v, l \cos u + l \log \tan \frac{u}{2}\Big).$$



Recordemos también que la longitud de arco es $s = -l \log \operatorname{sen} u$, luego $\operatorname{sen} u = e^{-s/l}$. La carta de P que resulta de tomar la tractriz parametrizada por el arco tiene la primera forma fundamental determinada por

$$E = 1, \quad F = 0, \quad G = r(s)^2 = l^2 e^{-2s/l}.$$

De aquí se sigue fácilmente que $K = -1/l^2$.

Por lo tanto un ejemplo de superficie de curvatura constante igual a $K < 0$ es la pseudoesfera

$$\frac{1}{\sqrt{-K}} \Big(\operatorname{sen} u \cos v, \operatorname{sen} u \operatorname{sen} v, \cos u + \log \tan \frac{u}{2}\Big).$$

■

***Nota** La pseudoesfera es al plano hiperbólico lo que un cilindro es al plano euclídeo. En efecto, hemos visto que si parametrizamos por el arco la tractriz obtenemos una carta de la pseudoesfera cuya primera forma fundamental es (para $l = 1$)

$$ds^2 = dw^2 + e^{-2u}\, dv^2,$$

donde $w \in \,]0, +\infty[$ es la longitud de arco de la tractriz (que arriba representábamos por s). Las cartas de la pseudoesfera tienen dominios de la forma $(w, v) \in \,]0, +\infty[\times]v_0 - \pi, v_0 + \pi[$. Si ahora hacemos el cambio $(w, v) = (\log y, x)$

obtenemos cartas con dominios de la forma $]x_0 - \pi, x_0 + \pi[\, \times\,]1, +\infty[$ de modo que la primera forma fundamental pasa a ser

$$ds^2 = \frac{dx^2 + dy^2}{y^2},$$

es decir, exactamente la del semiplano de Poincaré. Un cálculo rutinario nos da la forma explícita de estas cartas:

$$X(x,y) = \left(\frac{\cos x}{y}, \frac{\operatorname{sen} x}{y}, -\frac{\sqrt{y^2-1}}{y} + \frac{1}{2}\log(y-1) + \frac{1}{2}\log\big(y + \sqrt{y^2-1}\,\big).\right)$$

Esto significa que un fragmento del semiplano de Poincaré de la forma $]x_0 - \pi, x_0 + \pi[\, \times\,]1, +\infty[$ puede verse como un mapa de la pseudoesfera de modo que la longitud hiperbólica en el mapa coincide con la longitud euclídea sobre la superficie. Por lo tanto la porción de pseudoesfera cubierta por la carta (toda ella menos un meridiano $x =$cte.) puede identificarse con un fragmento de plano hiperbólico exactamente igual que una porción de esfera puede identificarse con un fragmento de plano elíptico.

La situación es, como decíamos, análoga a la del cilindro dado por

$$\big(r\cos(v/r), r\operatorname{sen}(v/r), u\big),$$

cuya primera forma fundamental es $ds^2 = dx^2 + dy^2$, igual que la del plano. La diferencia es que, en este caso, al quitarle una recta $x =$cte. podemos desplegarlo hasta hacerlo plano sin modificar su primera forma fundamental, mientras que la pseudoesfera no puede desplegarse sin sufrir estiramientos que alteren su métrica y su curvatura. Por ello no podemos extenderla a un plano hiperbólico completo. ■

Capítulo VI

Ecuaciones diferenciales ordinarias

Una *ecuación diferencial ordinaria* es una relación de la forma

$$f\big(t, y(t), y'(t), \ldots, y^{n)}(t)\big) = 0,$$

donde $f : D \subset \mathbb{R}^{n+2} \longrightarrow \mathbb{R}$ e $y : I \longrightarrow \mathbb{R}$ es una función definida en un intervalo I derivable n veces. Normalmente la función y es desconocida, y entonces se plantea el problema de *integrar* la ecuación, es decir, encontrar todas las funciones y que la satisfacen. El adjetivo "ordinaria" se usa para indicar que la función incógnita tiene una sola variable. Las ecuaciones que relacionan las derivadas parciales de una función de varias variables se llaman *ecuaciones diferenciales en derivadas parciales,* pero no vamos a ocuparnos de ellas. Dentro de las ecuaciones ordinarias, nos vamos a ocupar únicamente de un caso más simple pero suficientemente general: aquel en que tenemos despejada la derivada de orden mayor, es decir, una ecuación de la forma

$$y^{n)}(t) = f\big(t, y(t), y'(t), \ldots, y^{n-1)}(t)\big).$$

El número n se llama *orden* de la ecuación. El caso más simple es la ecuación de primer orden $y' = f(t)$. Sabemos que si la ecuación tiene solución de hecho hay infinitas de ellas pero, en un intervalo dado, cada una se diferencia de las demás en una constante, de modo que una solución queda completamente determinada cuando se especifica un valor $y(t_0) = y_0$. Veremos que esto sigue siendo válido para todas las ecuaciones de primer orden. Por ello se define un *problema de Cauchy* como

$$\left.\begin{array}{rl} y' &= f(t, y) \\ y(t_0) &= y_0 \end{array}\right\}$$

Resolver el problema significa encontrar una función y definida alrededor de t_0 de modo que satisfaga la ecuación diferencial y cumpla la condición inicial $y(t_0) = y_0$. Probaremos que bajo condiciones muy generales los problemas de Cauchy tienen solución única.

Toda la teoría se aplica igualmente al caso de sistemas de ecuaciones diferenciales. De hecho un sistema de ecuaciones puede verse como una única ecuación vectorial. Basta considerar que $y : I \longrightarrow \mathbb{R}^n$ y $f : D \subset \mathbb{R}^{n+1} \longrightarrow \mathbb{R}^n$.

Son muchas las ocasiones en las que el único conocimiento que tenemos de una o varias funciones es el hecho de que satisfacen un sistema de ecuaciones diferenciales. Por ejemplo, las ecuaciones (5.10) del capítulo anterior son un sistema de ecuaciones de segundo orden que determinan cuándo una curva $x(s)$ representa a una geodésica de una variedad en una carta dada. Los resultados que probaremos en este capítulo nos asegurarán en particular la existencia de geodésicas. De momento no tenemos garantizada la existencia de solución ni en el caso más simple: $y' = f(x)$. Éste es el primer punto que hemos de estudiar, lo que nos lleva a profundizar un poco más en el cálculo integral.

6.1 La integral de Riemann

Recordemos que hemos definido la expresión

$$\int_a^b f(x)\, dx$$

como $F(b) - F(a)$, donde F es una primitiva de f, pero la interpretación geométrica era el número que resulta de dividir el intervalo $[a, b]$ en intervalos infinitesimales de longitud dx y sumar los incrementos infinitesimales $f(x)\, dx$. Vamos a dar rigor a esta idea, lo que nos llevará a una construcción de la integral que no postule la existencia de la primitiva.

La técnica será, por supuesto, sustituir la división en infinitos intervalos infinitesimales por particiones en intervalos de longitud arbitrariamente pequeña. Puede probarse que no importa cómo escojamos estas particiones, por lo que trabajaremos concretamente con intervalos de longitud 2^{-n}.

Para cada número natural n sea $P_n = \{2^{-n}k \mid k \in \mathbb{Z}\}$. Para cada $x \in P_n$ sea $x'_n = x + 2^{-n}$. De este modo, la recta real se divide en una unión disjunta de intervalos de longitud 2^{-n}

$$\mathbb{R} = \bigcup_{x \in P_n} [x, x'_n[\,.$$

Llamaremos $\mathcal{F}$ al conjunto de todas las funciones $f : \mathbb{R} \longrightarrow \mathbb{R}$ tales que el conjunto $\{x \in \mathbb{R} \mid f(x) \neq 0\}$ está acotado. Para cada $f \in \mathcal{F}$ definimos

$$S_n(f) = \sum_{x \in P_n} f(x) 2^{-n}.$$

Notar que f se anula en todos los puntos de P_n salvo a lo sumo en un número finito de ellos, luego la suma anterior es en realidad una suma finita. La suma $S_n(f)$ es la aproximación de la integral de f que resulta de aproximar el incremento infinitesimal dx por el incremento finito 2^{-n}.

Diremos que f es *integrable Riemann* si existe

$$\int f(x)\,dx = \lim_n S_n(f) \in \mathbb{R}.$$

A esta cantidad la llamaremos *integral de Riemann* de f. Llamaremos $\mathcal{R}$ al conjunto de todas las funciones de $\mathcal{F}$ que son integrables Riemann.

Si $A \subset \mathbb{R}$, llamaremos *función característica* de A a la función $\chi_A : \mathbb{R} \longrightarrow \mathbb{R}$ dada por

$$\chi_A(x) = \begin{cases} 1 & \text{si } x \in A \\ 0 & \text{si } x \notin A \end{cases}$$

Dado un intervalo $[a, b]$ y una función $f : \mathbb{R} \longrightarrow \mathbb{R}$, es claro que $f\chi_{[a,b]} \in \mathcal{F}$. Diremos que f es integrable Riemann en $[a, b]$ si $f\chi_{[a,b]}$ es integrable Riemann. En tal caso llamaremos

$$\int_a^b f(x)\,dx = \int f(x)\chi_A(x)\,dx.$$

Es obvio que la integrabilidad de f en $[a, b]$ y, en su caso, el valor de la integral sólo dependen de la restricción de f al intervalo considerado. Por lo tanto, si f es una función definida en $[a, b]$, diremos que es integrable Riemann en $[a, b]$ si lo es la extensión a $\mathbb{R}$ que toma el valor 0 en todos los puntos exteriores a $[a, b]$. Así, una función es integrable en $[a, b]$ si y sólo si lo es, en este sentido, su restricción a dicho intervalo.

Llamaremos $\mathcal{R}(a, b)$ al conjunto de todas las funciones integrables Riemann en $[a, b]$.

Observemos que si $f \in \mathcal{R}$, entonces existe un intervalo $[a, b]$ tal que f se anula fuera de $[a, b]$, con lo que $f = f\chi_{[a,b]}$ y por lo tanto $f \in \mathcal{R}(a, b)$ y

$$\int f(x)\,dx = \int_a^b f(x)\,dx.$$

Por consiguiente no perdemos generalidad si trabajamos con funciones integrables en un intervalo fijo.

Teorema 6.1 *Sea $[a, b]$ un intervalo. Se cumplen las propiedades siguientes:*

a) Si f, $g \in \mathcal{R}(a, b)$ y α, $\beta \in \mathbb{R}$ entonces $\alpha f + \beta g \in \mathcal{R}(a, b)$ y

$$\int_a^b \big(\alpha f(x) + \beta g(x)\big)\,dx = \alpha \int_a^b f(x)\,dx + \beta \int_a^b f(x)\,dx.$$

b) Si f, $g \in \mathcal{R}(a, b)$ y $f \leq g$ entonces

$$\int_a^b f(x)\,dx \leq \int_a^b g(x)\,dx.$$

c) Si $f \in \mathcal{R}(a,b)$ y $|f|$ también es integrable entonces

$$\left| \int_a^b f(x)\, dx \right| \leq \int_a^b |f(x)|\, dx.$$

d) Si $k \in [a,b]$ entonces $\chi_{\{k\}} \in \mathcal{R}(a,b)$ y

$$\int_a^b \chi_{\{k\}}\, dx = 0.$$

e) Si $a < c < b$, $f \in \mathcal{R}(a,c)$ y $f \in \mathcal{R}(c,b)$ entonces $f \in \mathcal{R}(a,b)$ y

$$\int_a^b f(x)\, dx = \int_a^c f(x)\, dx + \int_c^b f(x)\, dx.$$

f) La función constante igual a 1 es integrable en $[a,b]$ y

$$\int_a^b dx = b - a.$$

Demostración: Las tres primeras propiedades son obvias para las sumas S_n y se trasladan a las integrales por las propiedades de los límites. Para probar d) observamos que si $k \in P_n$ para algún n_0 entonces $S_n(\chi_{\{k\}}) = 2^{-n}$, para $n \geq n_0$ y en caso contrario $S_n(\chi_{\{k\}}) = 0$ para todo n. En cualquier caso el límite cuanto n tiende a infinito vale 0.

Para probar e) observamos que

$$f\chi_{[a,b]} = f\chi_{[a,c]} + f\chi_{[c,b]} - f(c)\chi_{\{c\}}.$$

El resultado es inmediato a partir de los apartados anteriores. Para demostrar f) consideramos la suma

$$S_n(\chi_{[a,b]}) = \sum_{x \in P_n} \chi_{[a,b]}(x) 2^{-n}.$$

Si los puntos de P_n contenidos en $[a,b]$ son $a \leq x_0 < \cdots < x_k \leq b$, es claro que $k2^{-n} = x_k - x_0 \leq b - a$ y $(b-a) - k2^{-n} = x_0 - a + b - x_k \leq 2^{-n+1}$. Por lo tanto

$$|S_n(\chi_{[a,b]}) - (b-a)| = |(k+1)2^{-n} - (b-a)| \leq 2^{-n+1} + 2^{-n},$$

luego existe

$$\lim_n S_n(\chi_{[a,b]}) = b - a.$$

■

El teorema anterior puede mejorarse considerablemente. Por ejemplo, puede probarse que si f es integrable entonces $|f|$ también lo es, con lo que sobra la

hipótesis correspondiente en el apartado c). Así mismo, si $a < c < b$, la integrabilidad de f en $[a, b]$ implica la integrabilidad en $[a, c]$ y $[c, b]$. También puede probarse que el producto de funciones integrables es integrable. No entraremos en todo esto porque vamos a trabajar únicamente con funciones continuas, y estos hechos resultan triviales en este caso una vez probado el teorema siguiente.

Teorema 6.2 *Toda función continua en un intervalo* $[a, b]$ *es integrable Riemann en* $[a, b]$.

DEMOSTRACIÓN: En la prueba usaremos la siguiente observación sobre las sumas $S_n(f)$: Si una función f es constante sobre cada intervalo $[x, x'_n[$ con $x \in P_n$, entonces $S_n(f) = S_{n+1}(f)$.

En efecto, los puntos de P_{n+1} son los de P_n y los de la forma $x + 2^{-n-1}$, con $x \in P_n$. Por consiguiente

$$S_{n+1}(f) = \sum_{x \in P_n} \big(f(x) + f(x + 2^{-n-1})\big) 2^{-n-1} = \sum_{x \in P_n} 2f(x) 2^{-n-1} = S_n(f).$$

De aquí se sigue a su vez que $S_n(f) = S_m(f)$ para $m \geq n$.

Consideremos ahora una función f continua en $[a, b]$ y sea $\epsilon > 0$. Por el teorema 2.36 sabemos que f es uniformemente continua en $[a, b]$, luego existe un $\delta > 0$ tal que si x, $x' \in [a, b]$ cumplen $|x - x'| < \delta$ entonces

$$|f(x) - f(x')| < \frac{\epsilon}{2(b - a)}.$$

Sea n_0 un número natural tal que $1/n_0 < \delta$. Podemos exigir además que $2^{-n_0+1} < b - a$. Para cada $x \in P_{n_0}$ sea

$$\begin{aligned} m_x &= \text{ínf}\,\{f(t) \mid t \in \big[x, x'_{n_0}\big[\cap [a, b]\}, \\ M_x &= \sup\{f(t) \mid t \in \big[x, x'_{n_0}\big[\cap [a, b]\}. \end{aligned}$$

Estos supremos e ínfimos existen porque f está acotada en $[a, b]$ (por compacidad). Entendemos que si $\big[x, x'_{n_0}\big[\cap [a, b] = \varnothing$ entonces $m_x = M_x = 0$.

Puesto que $x'_{n_0} - x < 2^{-n_0} < \delta$ tenemos que la distancia entre dos valores de f en un intervalo $\big[x, x'_{n_0}\big[\cap [a, b]$ no es superior a $\epsilon/2(b - a)$, de donde se concluye inmediatamente que $M_x - m_x \leq \epsilon/2(b - a)$.

Llamemos g y h a las funciones que en cada intervalo $\big[x, x'_{n_0}\big[$ toman el valor constante m_x y M_x respectivamente. Entonces es claro que $g \leq f \leq h$ y $0 \leq h - g \leq \epsilon/2(b - a)$.

Para todo $n \geq n_0$ se cumple

$$S_{n_0}(g) = S_n(g) \leq S_n(f) \leq S_n(h) = S_{n_0}(h),$$

luego para $n, m \geq n_0$ se cumple

$$|S_n(f) - S_m(f)| \leq S_{n_0}(h - g) = \sum_{x \in P_{n_0}} (M_x - m_x) 2^{-n_0}$$

Si llamamos k al número de puntos $x \in P_{n_0}$ tales que $\big[x, x'_{n_0}\big[\subset [a,b]$ es claro que $k2^{-n_0} \leq b-a$. Puede ocurrir que haya dos puntos adicionales $y \in P_n$ tal que $\big[y, y'_{n_0}\big[$ no esté contenido en $[a,b]$ pero corte a $[a,b]$. En cualquier caso hay a lo sumo $k+2$ intervalos que cortan a $[a,b]$ y por lo tanto la suma anterior tiene a lo sumo $k+2$ sumandos no nulos. Así pues

$$|S_n(f) - S_m(f)| \leq (k+2)2^{-n_0}\,\frac{\epsilon}{2(b-a)} < \epsilon.$$

Esto prueba que la sucesión $S_n(f)$ es de Cauchy, luego converge, luego f es integrable Riemann. ■

Conviene introducir el convenio de que

$$\int_a^b f(x)\,dx = -\int_b^a f(x)\,dx, \qquad \int_a^a f(x)\,dx = 0.$$

De este modo la fórmula del apartado e) del teorema 6.1 se cumple para tres puntos cualesquiera a, b, c independientemente de cómo estén ordenados o de si son iguales o no. En particular, si f es una función continua en un intervalo I (no necesariamente acotado) y $a \in I$ podemos definir

$$F(x) = \int_a^x f(t)\,dt$$

para todo $x \in I$. Ahora probamos que la integral de Riemann de una función continua coincide con la integral calculada mediante una primitiva.

Teorema 6.3 *Sea f una función continua en un intervalo I y sea $a \in I$. Entonces la función*

$$F(x) = \int_a^x f(t)\,dt$$

es derivable en el interior de I y $F' = f$.

Demostración: Sea x un punto interior de I y sea $J \subset I$ un intervalo cerrado y acotado que contenga a a y a x (a éste último en su interior). Por el teorema 2.36 la función f es uniformemente continua en J, luego para cada $\epsilon > 0$ existe un $\delta > 0$ tal que si u, $u' \in J$, $|u-u'| < \delta$ entonces $|f(u)-f(u')| < \epsilon$.

Sea $h \in \mathbb{R}$ tal que $|h| < \delta$ y $x+h \in J$. Sean m y M el mínimo y el máximo de f en el intervalo cerrado de extremos x y $x+h$. Si $h > 0$

$$mh = \int_x^{x+h} m\,dt \leq \int_x^{x+h} f(t)\,dt \leq \int_x^{x+h} M\,dt = Mh.$$

Si $h < 0$ se invierten las desigualdades, pero en ambos casos resulta

$$m \leq \frac{\int_x^{x+h} f(t)\,dt}{h} \leq M.$$

Por el teorema de los valores intermedios existe un α entre x y $x+h$ de modo que

$$f(\alpha) = \frac{\int_x^{x+h} f(t)\,dt}{h} = \frac{F(x+h)-F(x)}{h}.$$

Claramente $|\alpha - x| < |h| < \delta$, luego

$$\left|\frac{F(x+h)-F(x)}{h} - f(x)\right| = |f(\alpha) - f(x)| < \epsilon,$$

por lo que existe $F'(x) = f(x)$. ■

Como consecuencia obtenemos:

Teorema 6.4 (Regla de Barrow) *Si f es una función continua en un intervalo $[a,b]$ entonces F tiene una primitiva F en $[a,b]$ y*

$$\int_a^b f(x)\,dx = F(b) - F(a).$$

Demostración: Cuando decimos que F es una primitiva de f en $[a,b]$ queremos decir que F es continua en $[a,b]$, derivable en $]a,b[$ y $F' = f$ en $]a,b[$. Basta tomar como F la función

$$F(x) = \int_c^x f(t)\,dt,$$

donde $a < c < b$. Sólo falta probar que F es continua en a y en b, lo cual es sencillo: Sea $|f(x)| \leq M$ en $[a,b]$. Entonces

$$|F(b) - F(x)| \leq \int_x^b |f(t)|\,dt \leq \int_x^b M\,dt = M(b-x),$$

luego si $|b-x| < \delta = \epsilon/M$ se cumple $|F(b) - F(x)| < \epsilon$, lo que prueba la continuidad en b, e igualmente sucede con a. ■

Puesto que toda función continua tiene primitiva, todo arco x de clase C^1 es rectificable, pues la función $\|x'\|$ es continua.

Ejemplo Vamos a demostrar el resultado que dejamos pendiente en el capítulo III, a saber, la convergencia de la serie de Taylor del arco tangente incluso en los puntos frontera ± 1. Para ello partimos de la suma parcial de la serie geométrica de razón $-t^2$:

$$\frac{1}{1+t^2} = 1 - t^2 + t^4 + \cdots + (-1)^n t^{2n} + (-1)^{n+1}\frac{t^{2n+2}}{1+t^2}.$$

Integrando ambos miembros resulta

$$\begin{aligned}\arctan x = \int_0^x \frac{1}{1+t^2}\,dt &= x - \frac{x^3}{3} + \frac{x^5}{5} - \cdots + (-1)^n \frac{x^{2n+1}}{2n+1} \\ &+ (-1)^{n+1} \int_0^x \frac{t^{2n+2}}{1+t^2}\,dt.\end{aligned}$$

El polinomio de la derecha es el polinomio de Taylor del arco tangente, luego

$$|R_{2n+1}(x)| = \left|\int_0^x \frac{t^{2n+2}}{1+t^2}\,dt\right| \leq \left|\int_0^x t^{2n+2}\,dt\right| = \frac{|x|^{2n+3}}{2n+3},$$

de donde se sigue que el resto tiende a cero incluso si $x = \pm 1$, como había que probar. ■

Finalmente definimos la integral de una función $f : [a, b] \longrightarrow \mathbb{R}^n$ como

$$\int_a^b f(x)\,dx = \left(\int_a^b f_1(x)\,dx, \ldots, \int_a^b f_n(x)\,dx\right),$$

entendiendo que f es integrable Riemann en $[a, b]$ si y sólo si lo son todas sus funciones coordenadas f_i.

6.2 Ecuaciones diferenciales de primer orden

Nos ocupamos ahora de asegurar la existencia y unicidad de la solución de los problemas de Cauchy. Nos basaremos en un resultado general sobre espacios métricos completos:

Teorema 6.5 (Teorema de punto fijo de Banach) *Sea M un espacio métrico completo y $T : M \longrightarrow M$ una aplicación tal que existe un número real $0 < \alpha < 1$ de modo que*

$$d\big(T(x), T(y)\big) < \alpha\, d(x, y), \quad \textit{para todo } x, y \in M.$$

Entonces existe un único $x \in M$ tal que $T(x) = x$.

Las aplicaciones T que cumplen la propiedad indicada se llaman *contractivas.* Los puntos x que cumplen $T(x) = x$ se llaman *puntos fijos* de T. El teorema afirma, pues, que toda aplicación contractiva en un espacio métrico completo tiene un único punto fijo.

DEMOSTRACIÓN: Tomamos un punto arbitrario $x_0 \in M$ y consideramos la sucesión dada por $x_{n+1} = T(x_n)$. Por la propiedad de T, tenemos que

$$\begin{aligned} d(x_1, x_2) &= d\big(T(x_0), T(x_1)\big) < \alpha\, d(x_0, x_1), \\ d(x_2, x_3) &= d\big(T(x_1), T(x_2)\big) < \alpha\, d(x_1, x_2) = \alpha^2\, d(x_0, x_1), \end{aligned}$$

y en general concluimos $d(x_n, x_{n+1}) < \alpha^n\, d(x_0, x_1)$. Aplicando la desigualdad triangular resulta, para $n < m$,

$$d(x_n, x_m) < \left(\sum_{i=n}^{m-1} \alpha^i\right) d(x_0, x_1) < \left(\sum_{i=n}^{\infty} \alpha^i\right) d(x_0, x_1) = \frac{\alpha^n}{1-\alpha}\, d(x_0, x_1).$$

El término de la derecha tiende a 0, lo que significa que la sucesión x_n es de Cauchy. Como el espacio M es completo existe $x = \lim_n x_n \in M$. Veamos que x es un punto fijo de T. Para ello observamos que

$$\begin{aligned} d\big(x, T(x)\big) &\leq d(x, x_n) + d(x_n, x_{n+1}) + d\big(x_{n+1}, T(x)\big) \\ &< (1+\alpha)d(x, x_n) + \alpha^n\, d(x_0, x_1). \end{aligned}$$

El último término tiende a 0, luego ha de ser $d\big(x, T(x)\big) = 0$, es decir, $T(x) = x$. Si y es otro punto fijo de T, entonces $d\big(T(x), T(y)\big) = d(x, y)$, en contradicción con la propiedad contractiva, luego el punto fijo es único. ■

Es frecuente que una ecuación diferencial dependa de uno más parámetros. Por ejemplo, la fuerza $F(t, x)$ que afecta a un móvil de masa m es, por lo general, función del tiempo t y de la posición x. La segunda ley de Newton afirma que su trayectoria $x(t)$ obedece la ecuación diferencial de segundo orden

$$F(t, x) = m\, x''(t),$$

donde la masa m es un parámetro. En casos como este podemos considerar la solución como función de los parámetros, es decir, $x(t, m)$ es la posición en el instante t de un cuerpo de masa m sometido a la fuerza $F(t, x)$ (y en unas condiciones iniciales dadas). En el teorema de existencia y unicidad que damos a continuación contemplamos la existencia de estos parámetros y probamos que la solución depende continuamente de ellos.

Teorema 6.6 *Sean t_0, a, $b_1, \ldots, b_n$, $y_1^0, \ldots, y_n^0$ números reales. Consideremos una aplicación continua*

$$f : [t_0 - a, t_0 + a] \times \prod_{i=1}^{n} [y_i^0 - b_i, y_i^0 + b_i] \times K \longrightarrow \mathbb{R}^n,$$

donde K es un espacio métrico compacto. Sea M una cota de f respecto a la norma $\|\ \|_\infty$ en $\mathbb{R}^n$. Supongamos que existe una constante N tal que

$$\|f(t, y, \mu) - f(t, z, \mu)\|_\infty \leq N\|y - z\|_\infty.$$

Entonces el problema de Cauchy

$$\left.\begin{aligned} y'(t, \mu) &= f(t, y, \mu) \\ y(t_0, \mu) &= y_0 \end{aligned}\right\}$$

tiene solución única $y : [t_0 - h_0, t_0 + h_0] \times K \longrightarrow \mathbb{R}^n$, continua en su dominio, donde h_0 es cualquier número real tal que

$$0 < h_0 \leq \text{mín}\left\{a, \frac{b_1}{M}, \ldots, \frac{b_n}{M}\right\}, \quad h_0 < \frac{1}{N}.$$

Se entiende que derivada de y que aparece en el problema de Cauchy es respecto de la variable t. Teóricamente deberíamos usar la notación de derivadas parciales, pero es costumbre usar la notación del análisis de una variable para evitar que el problema parezca una ecuación diferencial en derivadas parciales, cuando en realidad no lo es.

Demostración: Sea h_0 en las condiciones indicadas, sea $I = [t-h_0, t+h_0]$, sea $D = \prod_{i=1}^{n} [y_i^0 - b_i, y_i^0 + b_i]$ y sea $M = C(I \times K, D)$, que es un espacio de Banach con la norma supremo. Definimos el operador $T : M \longrightarrow M$ mediante

$$T(y)(t, \mu) = y^0 + \int_{t_0}^{t} f\big(t, y(t, \mu), \mu\big)\, dt.$$

Hemos de probar que $T(y)(t, \mu) \in D$ y que $T(y)$ es una aplicación continua. En primer lugar,

$$\begin{aligned} \big|T(y)_i(t,\mu) - y_i^0\big| &= \left|\int_{t_0}^{t} f_i(t,y,\mu)\, dt\right| \leq \left|\int_{t_0}^{t} |f_i(t,y,\mu)|\, dt\right| \leq M \left|\int_{t_0}^{t} dt\right| \\ &= M|t-t_0| \leq M h_0 \leq b_i. \end{aligned}$$

Esto prueba que $T(y)(t, \mu) \in D$. La continuidad es consecuencia de un cálculo rutinario:

$$\|T(y)(t_1,\mu_1) - T(y)(t_2,\mu_2)\|_\infty = \max_i \left| \int_{t_0}^{t_1} f_i(t,y,\mu_1)\, dt - \int_{t_0}^{t_2} f_i(t,y,\mu_2)\, dt \right|$$

$$\leq \max_i \left(\left| \int_{t_0}^{t_1} f_i(t,y,\mu_1)\, dt - \int_{t_0}^{t_1} f_i(t,y,\mu_2)\, dt \right| + \right.$$

$$\left. \left| \int_{t_0}^{t_1} f_i(t,y,\mu_2)\, dt - \int_{t_0}^{t_2} f_i(t,y,\mu_2)\, dt \right| \right)$$

$$\leq \max_i \left(\left| \int_{t_0}^{t_1} \big(f_i(t,y,\mu_1) - f_i(t,y,\mu_2)\big)\, dt \right| + \left| \int_{t_2}^{t_1} f_i(t,y,\mu_2)\, dt \right| \right)$$

$$\leq \max_i \left| \int_{t_0}^{t_1} \big|f_i(t,y,\mu_1) - f_i(t,y,\mu_2)\big|\, dt \right| + M|t_1 - t_2|.$$

Sea $\epsilon > 0$. La función $f_i(t, y(t, \mu), \mu)$ es uniformemente continua en el compacto $I \times K$, luego existe un $\delta > 0$ tal que si $d(\mu_1, \mu_2) < \delta$, entonces $|f_i(t, y, \mu_1) - f_i(t, y, \mu_2)| < \epsilon/2h_0$. Podemos suponer que esto vale para todo $i = 1, \ldots, n$, y si suponemos también que $|t_1 - t_2| < \epsilon/2M$ concluimos que

$$\|T(y)(t_1,\mu_1) - T(y)(t_2,\mu_2)\|_\infty < \epsilon.$$

Esto prueba la continuidad de $T(y)$ en el punto (t_1, μ_1).

Ahora probamos que T es contractivo, con constante $\alpha = Nh_0 < 1$. En efecto,

$$\|T(y)(t,\mu) - T(z)(t,\mu)\|_\infty = \max_i \left| \int_{t_0}^t \big(f_i(t, y(t,\mu),\mu) - f_i(t, z(t,\mu),\mu)\big)\, dt \right|$$

$$\leq \max_i \left| \int_{t_0}^t N\|y(t,\mu) - z(t,\mu)\|_\infty \, dt \right| \leq N \max_i \left| \int_{t_0}^t \|y - z\| \, dt \right| \leq Nh_0\|y - z\|.$$

Por definición de norma supremo resulta $\|T(y) - T(z)\| \leq \alpha\|y - z\|$. El teorema anterior implica ahora la existencia de una única función $y \in M$ tal que

$$y(t,\mu) = y^0 + \int_{t_0}^t f(t, y, \mu)\, dt,$$

pero es claro que esto equivale a ser solución del problema de Cauchy, luego éste tiene solución única. ■

En la práctica, hay una hipótesis más fuerte que la condición de Lipschitz que hemos exigido en el teorema anterior pero que es más fácil de comprobar. Se trata de exigir simplemente que la función f sea de clase C^1.

Teorema 6.7 *Sea $f : D \subset \mathbb{R}^n \longrightarrow \mathbb{R}^m$ una función de clase C^1 en un abierto D. Para todo subconjunto compacto convexo $C \subset D$ existe una constante N tal que si y, $z \in C$ entonces $\|f(y) - f(z)\|_\infty \leq N\|y - z\|_\infty$.*

Demostración: Si llamamos $f_1, \ldots, f_m$ a las funciones coordenadas de f, basta probar que $|f_i(y) - f_i(z)| \leq N_i\|y - z\|_\infty$ para todo y, $z \in C$, pues tomando como N la mayor de las constantes N_i se cumple la desigualdad buscada. Equivalentemente, podemos suponer que $m = 1$.

Dados y, $z \in C$, consideramos la función $g(h) = f\big(y + h(z - y)\big)$, definida en $[0, 1]$, pues C es convexo. Se cumple $g(0) = f(y)$, $g(1) = f(z)$. Por el teorema del valor medio existe $0 < h_0 < 1$ tal que

$$f(z) - f(y) = g'(h_0) = df(\xi)(z - y) = \nabla f(\xi)(z - y),$$

donde $\xi = y + h_0(z - y) \in C$. Sea N_0 una cota del módulo de las derivadas parciales de f (que por hipótesis son continuas) en el compacto C. Entonces

$$|f(z) - f(y)| = \left| \sum_{i=1}^n D_i f(\xi)(z_i - y_i) \right| \leq \sum_{i=1}^n N_0\|z - y\|_\infty = nN_0\|z - y\|_\infty.$$

■

En realidad, si tomamos como hipótesis que la ecuación diferencial sea de clase C^1 obtenemos no sólo la continuidad de la solución respecto de los parámetros, sino también la derivabilidad.

Teorema 6.8 *Sea* $f : D \subset \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ *una función de clase* C^1 *en el abierto* D. *Sea* $(t_0, y_0, \mu) \in D$. *Entonces el problema de Cauchy*

$$\left.\begin{aligned} y'(t,\mu) &= f(t,y,\mu) \\ y(t_0,\mu) &= y_0 \end{aligned}\right\}$$

tiene solución única definida y de clase C^1 *en un entorno de* (t_0, μ).

Demostración: Tomamos un entorno C de (t_0, y_0, μ) que esté contenido en D y sea producto de intervalos cerrados de centro cada una de las componentes del punto. En particular es convexo y compacto. El teorema anterior garantiza que se cumplen las hipótesis del teorema de existencia y unicidad. Falta probar que la función $y(t,\mu)$ es de clase C^1 en su dominio.

Obviamente y es derivable respecto de t y la derivada es continua. Veamos que lo mismo sucede con las demás variables. Sea e_i un vector de la base canónica de $\mathbb{R}^m$. Consideramos un punto (t_1, μ_1) del dominio de y. La función

$$Q(t,\mu,h) = \frac{y(t,\mu+he_i) - y(t,\mu)}{h}$$

está definida en los puntos de un entorno de $(t_1, \mu_1, 0)$ para los que $h \neq 0$. Hemos de probar que tiene límite cuando h tiende a 0. Claramente

$$\frac{\partial Q}{\partial t} = \frac{f(t, y(t,\mu+he_i), \mu+he_i) - f(t, y(t,\mu), \mu)}{h}.$$

Llamemos

$$E(p,x) = \begin{cases} \dfrac{f(p+x) - f(p) - df(p)(x)}{\|x\|} & \text{si } x \neq 0 \\ 0 & \text{si } x = 0 \end{cases}$$

Se comprueba que es continua en un entorno de $(t_1, y(t_1,\mu_1), \mu_1, 0)$. Sólo hay que ver la continuidad en los puntos de la forma $(q, 0)$. Se demuestra para cada función coordenada independientemente, y a su vez para ello se aplica el teorema del valor medio la la función $f_j(p+tx)$. El resultado es que

$$E_j(p,x) = \big(\nabla f_j(p') - \nabla f_j(p)\big)\frac{x}{\|x\|},$$

donde p' es un punto entre p y $p+x$, y ahora basta aplicar la continuidad de las derivadas parciales de f.

En términos de E tenemos

$$\frac{\partial Q}{\partial t} = df(t, y(t,\mu), \mu)\big(0, Q(t,\mu,h), e_i\big)$$

$$+\|(0, Q(t,\mu,h), 1)\|\frac{|h|}{h}\, E(0, y(t,\mu+he_i) - y(t,\mu), h).$$

Más brevemente

$$\frac{\partial Q}{\partial t} = df(t, y(t,\mu), \mu)\big(0, Q(t,\mu,h), e_i\big) + \|(0, Q(t,\mu,h), 1)\|\, E^*(t,\mu,h),$$

entendiendo que $E^*(t,\mu,h)$ es continua en un entorno de $(t_1, \mu_1, 0)$ y se anula en los puntos donde $h = 0$.

Esto significa que Q es la solución de una ecuación diferencial determinada por la función continua

$$g(t, Q, \mu, h) = df(t, y(t,\mu), \mu)\big(0, Q, e_i\big) + \|(0, Q, 1)\|\, E^*(t,\mu,h),$$

donde μ y h son parámetros. Esta función no es diferenciable, pero cumple claramente la hipótesis del teorema 6.6. Concretamente la consideramos definida en un producto de intervalos de centros t_1, $Q(t_1, \mu_1, h)$ (para un h fijo) por un entorno compacto K de (μ_1, h) que contenga a $(\mu_1, 0)$. Si tomamos como condición inicial en el punto (t_1, μ_1, h) la determinada por la función Q que ya tenemos definida, el teorema 6.6 nos garantiza la existencia de solución continua en un conjunto de la forma $[t_1 - r, t_1 + r] \times K$. Por la unicidad la solución debe coincidir con la función Q que ya teníamos. En particular coincidirá con ella en los puntos de un entorno de $(t_1, \mu_1, 0)$ tales que $h \neq 0$. De aquí se sigue que existe

$$\lim_{h \to 0} Q(t_1, \mu_1, h) = \frac{\partial y}{\partial \mu_i}(t_1, \mu_1).$$

Más aún, esta derivada satisface la ecuación diferencial

$$\frac{\partial}{\partial t}\frac{\partial y}{\partial \mu_i} = df\big(t, y(t,\mu), \mu\big)\left(0, \frac{\partial y}{\partial \mu_i}, e_i\right).$$

Esto implica que es continua, luego y es de clase C^1. ∎

Hemos probado que las derivadas respecto a los parámetros de una ecuación diferencial determinada por una función de clase C^k satisfacen una ecuación diferencial de clase C^{k-1}. Una simple inducción prueba entonces que la solución y de una ecuación de clase C^k es una función de clase C^k.

Notemos que la solución y de un problema de Cauchy puede considerarse también como función de las condiciones iniciales, es decir, $y(t, \mu, t_0, y_0)$. Del teorema anterior se deduce que y es continua respecto a todas las variables, es decir, como función definida en un entorno de (t_0, μ, t_0, y_0) en $\mathbb{R} \times \mathbb{R}^m \times \mathbb{R} \times \mathbb{R}^n$. Para ello basta ver que si hacemos $z = y(t, \mu, t_0, y_0) - y_0$ y $r = t - t_0$, el problema

$$\left.\begin{aligned} y'(t) &= f(t, y, \mu) \\ y(t_0) &= y_0 \end{aligned}\right\}$$

es equivalente a

$$\left.\begin{aligned} z'(r) &= f(r + t_0, z + y_0, \mu) \\ z(0) &= 0 \end{aligned}\right\}$$

en el sentido de que una solución de uno da una del otro mediante los cambios de variable indicados. El segundo miembro del segundo problema es una función

$g(r,z,\nu)$, donde $\nu = (t_0, y_0, \mu) \in D$. Concretamente, el dominio de g es el abierto

$$\{(r, z, t_0, y_0, \mu) \mid (r + t_0, z + y_0, \mu) \in D,\ (t_0, y_0, \mu) \in D\}.$$

Una solución z del segundo problema definida en un entorno de $(0, t_0, y_0, \mu)$ se traduce en una solución $y(t, \mu, t_0, y_0)$ del primer problema definida en un entorno de (t_0, μ, t_0, y_0). En definitiva tenemos el siguiente enunciado, más completo, del teorema de existencia y unicidad:

Teorema 6.9 *Sea $f : D \subset \mathbb{R} \times \mathbb{R}^n \times \mathbb{R}^m \longrightarrow \mathbb{R}^n$ una función de clase C^k ($k \geq 1$) en el abierto D. Sea $(t_0, y_0, \mu) \in D$. Entonces el problema de Cauchy*

$$\left.\begin{aligned} y'(t, \mu, t_0, y_0) &= f(t, y, \mu) \\ y(t_0, \mu, t_0, y_0) &= y_0 \end{aligned}\right\}$$

tiene solución única de clase C^k en un entorno de (t_0, μ, t_0, y_0).

Conviene interpretar las ecuaciones diferenciales de primer orden en términos cercanos a las aplicaciones físicas. Supongamos que D es una región del espacio ocupada por un fluido en movimiento, como puede ser el aire o un caudal de agua. Podemos considerar entonces la función $V : I \times D \subset \mathbb{R} \times \mathbb{R}^n \longrightarrow \mathbb{R}^n$ que determina la velocidad del fluido $V_t(x)$ en cada instante $t \in I$ y en cada punto $x \in D$. Es lo que se llama un *campo de velocidades variable.* Si V no depende de t tenemos un *campo de velocidades estacionario.* Entonces, la solución del problema de Cauchy

$$\left.\begin{aligned} x'(t) &= V_t\big(x(t)\big) \\ x(t_0) &= x_0 \end{aligned}\right\}$$

se interpreta como la trayectoria que seguirá un cuerpo de masa despreciable (un papel en el aire) abandonado en el punto x_0 en el instante t_0. Las soluciones se llaman *líneas de flujo* del campo de velocidades. Estos problemas son el objeto de estudio de la *hidrodinámica,* o *mecánica de fluidos.* Observemos que cualquier problema de Cauchy puede interpretarse de este modo, lo cual es conveniente en muchas ocasiones. La función $x(t, \mu, t_0, y_0)$ dada por el teorema anterior se conoce como *flujo* del campo de velocidades. Cuando el campo es estacionario el instante inicial t_0 es irrelevante, pues claramente $x(t, \mu, t_0, y_0) = x(t - t_0, \mu, 0, y_0)$, por lo que podemos suponer siempre que $t_0 = 0$ y entender que el flujo $x(t, \mu, y_0)$ indica la posición de un objeto sometido al campo t unidades de tiempo después de que se encontrara en la posición y_0.

Otra cuestión importante es el dominio de la solución $y(t)$ (para unos parámetros y condiciones iniciales dados). En principio hemos probado que la ecuación tiene solución única en un entorno de t_0, pero una solución dada en un entorno dado puede admitir prolongaciones a un intervalo mayor. Haciendo uso de las estimaciones explícitas del teorema 6.6 junto con la unicidad de las soluciones es fácil ver que dos prolongaciones de una misma solución han de coincidir en su dominio común, luego la unión de todas las prolongaciones constituye una solución máxima, no prolongable, de la ecuación diferencial. El dominio de esta solución máxima ha de ser un intervalo abierto, sin que pueda prolongarse por

continuidad a sus extremos, o de lo contrario podríamos usar el teorema de existencia para prolongar la solución un poco más tomando como condiciones iniciales los valores en dicho extremo. Si el dominio de la solución máxima está acotado superior o inferiormente, ello se debe necesariamente a que la curva obtenida tiende a infinito en el extremo o bien se acerca a la frontera del abierto donde está definido el problema. No entraremos en detalles sobre esto.

Como ejemplo de aplicación del teorema anterior demostramos lo siguiente:

Teorema 6.10 *Dadas dos funciones* $\kappa, \tau : I \longrightarrow \mathbb{R}$ *de clase* C^2 *de modo que* $\kappa \geq 0$ *y un punto* $t_0 \in I$, *existe una curva regular* $x : \,]t_0 - \epsilon, t_0 + \epsilon[\, \longrightarrow \mathbb{R}^3$ *parametrizada por el arco tal que* κ *y* τ *son respectivamente su curvatura y su torsión. La curva es única salvo isometrías.*

Demostración: La unicidad nos la da el teorema 4.23. Para probar la existencia consideramos el sistema de ecuaciones diferenciales determinado por las fórmulas de Frenet:

$$T' = \kappa N, \quad N' = -\kappa T - \tau B, \quad B' = \tau N.$$

Se trata de un sistema de nueve ecuaciones diferenciales con incógnitas las nueve funciones coordenadas de T, N y B. Tomamos unas condiciones iniciales cualesquiera T_0, N_0, B_0 tales que formen una base ortonormal positivamente orientada. Por el teorema anterior existen unas únicas funciones (T, B, N) que satisfacen las ecuaciones. Un simple cálculo nos da

$$\begin{array}{rclcrcl} (TN)' &=& \kappa NN - \kappa TT - \tau TB, & & (TB)' &=& \kappa NB + \tau TN, \\ (NB)' &=& -\kappa TB - \tau BB + \tau NN & & (TT)' &=& 2\kappa TN, \\ (NN)' &=& -2\kappa TN - 2\tau NB & & (BB)' &=& 2\tau NB. \end{array}$$

Vemos que las seis funciones TN, TB, NB, TT, NN, BB satisfacen un sistema de ecuaciones diferenciales con la condición inicial $(0, 0, 0, 1, 1, 1)$ que por otra parte es claro que tiene por solución a la función constante $(0, 0, 0, 1, 1, 1)$. La unicidad implica que (T, N, B) es una base ortonormal de $\mathbb{R}^3$.

Definimos

$$x(s) = \int_{t_0}^{s} T(s)\, ds.$$

Entonces es claro que $x'(s) = T(s)$, luego en particular x está parametrizada por el arco. Además $x''(s) = \kappa N$, luego κ es la curvatura de x. Un simple cálculo nos da que la torsión es τ. ■

Veamos ahora una aplicación a las variedades diferenciables. Sea $S \subset \mathbb{R}^m$ una variedad de dimensión n y $\alpha : I \longrightarrow S$ una curva parametrizada por el arco. Sea $\alpha(s_0) = p$. Sea X una carta de S alrededor de p. Entonces α tiene asociada una representación en la carta $x(s)$ de modo que $\alpha(s) = X\big(x(s)\big)$. Un vector arbitrario $w_0 \in T_p(S)$ es de la forma $w_0 = dX(x(s_0))(a_0)$. Teniendo en cuenta las ecuaciones (5.8) del capítulo anterior es claro que la solución $a(s)$ del problema

$$a_k' + \sum_{i,j=1}^{n} a_i \Gamma_{ij}^k x_j' = 0$$

con la condición inicial $a(s_0) = a_0$ determina un campo de vectores

$$w(s) = a_1(s)\, D_1X\big(x(s)\big) + \cdots + a_n(s)\, D_nX\big(x(s)\big)$$

con derivada covariante nula. Con esto casi tenemos demostrado el teorema siguiente:

Teorema *Sea S una variedad y $\alpha : I \longrightarrow S$ una curva parametrizada por el arco. Sea $\alpha(s_0) = p$ y sea $w_0 \in T_p(S)$. Entonces existe un único campo $w : I \longrightarrow \mathbb{R}^3$ tal que $w(s) \in T_{\alpha(s)}(S)$, $w(s_0) = w_0$ y $Dw = 0$. Lo llamaremos* transporte paralelo *de w_0 a través de α.*

En realidad hemos probado la existencia de w en un entorno de s_0. Vamos a justificar que la solución puede prolongarse a todo I. Para ello conviene observar que al ser $Dw = 0$ tenemos que $w'(s)$ es perpendicular a $T_{\alpha(s)}(S)$, luego $(ww)' = 2ww' = 0$, es decir, $\|w\|$ es constante. De aquí se sigue que el transporte paralelo es único: si w y w' son transportes paralelos de un mismo vector, entonces $w - w'$ es un transporte paralelo del vector nulo (porque su derivada covariante es la resta de las de los dos campos, luego es nula), luego es la aplicación constantemente nula.

Acabamos de usar la linealidad de la derivada covariante. Más en general, si tenemos dos vectores w_0, $w_0' \in T_p(S)$ y ambos tienen transporte paralelo w y w', entonces $\alpha w + \beta w'$ es un transporte paralelo de $\alpha w_0 + \beta w_0'$. Según lo que sabemos, existe un entorno de s_0 a lo largo del cual todos los vectores de una base de $T_p(S)$ tienen transporte paralelo, con lo que de hecho todos los vectores de $T_p(S)$ lo tienen.

Dado cualquier $s \in I$, es claro que el intervalo $[s_0, s]$ (o $[s, s_0]$ si $s < s_0$) puede cubrirse con un número finito de estos entornos donde existe transporte paralelo, de donde se sigue inmediatamente la existencia de transporte paralelo desde s_0 hasta s, luego el transporte paralelo existe sobre todo I. ■

Definición 6.11 Sea S una variedad y $\alpha : [s_0, s_1] \longrightarrow S$ una curva parametrizada por el arco de extremos p y q. Según el teorema anterior, para cada vector $w_0 \in T_p(S)$ tenemos definido el transporte paralelo $w(s)$ a lo largo de α de modo que $w(s_0) = w_0$. Al vector $\mathrm{tp}_{pq}^{\alpha}(w_0) = w(s_1) \in T_q(S)$ lo llamaremos *trasladado* de w_0 a lo largo de α. Esto nos define una aplicación $\mathrm{tp}_{pq}^{\alpha} : T_p(S) \longrightarrow T_q(S)$ a la que llamaremos *transporte paralelo* de $T_p(S)$ a $T_q(S)$ a lo largo de α.

Ejercicio: Probar que el transporte paralelo $\mathrm{tp}_{pq}^{\alpha} : T_p(S) \longrightarrow T_q(S)$ es una isometría.

6.3 Ecuaciones diferenciales de orden superior

Las ecuaciones diferenciales que aparecen con mayor frecuencia en física y en geometría son de orden 2. Afortunadamente, toda la teoría sobre existencia y unicidad que vamos a necesitar para ecuaciones diferenciales de orden superior se deduce inmediatamente del caso de orden 1. En efecto:

Teorema 6.12 *Sea $f : D \subset \mathbb{R} \times \mathbb{R}^{nm} \times \mathbb{R}^k \longrightarrow \mathbb{R}^n$ una función de clase C^k con $k \geq 1$ en un abierto D. Entonces la ecuación diferencial*

$$\left.\begin{array}{rcl} y^{m)}(t) &=& f(t, y, y', \ldots, y^{m-1)}, \mu) \\ y(t_0) &=& y_0 \\ y'(t_0) &=& y'_0 \\ \cdots\cdots && \cdots \\ y^{m-1)}(t_0) &=& y_0^{m-1)} \end{array}\right\}$$

tiene solución única $y(t, \mu, t_0, y_0, y'_0, \ldots, y_0^{m-1)})$ de clase C^k en un entorno de cada punto $(t_0, \mu, t_0, y_0, y'_0, \ldots, y_0^{m-1)})$.

Demostración: Basta observar que el problema equivale al sistema de ecuaciones de primer orden

$$\left.\begin{array}{rcl} y'(t) &=& y_1 \\ y'_1(t) &=& y_2 \\ \cdots\cdots && \cdots \\ y'_{m-2}(t) &=& y_{m-1} \\ y'_{m-1}(t) &=& f(t, y, y_1, \ldots, y_{m-1}, \mu) \\ (y, y_1, \ldots, y_{m-1})(t_0) &=& (y_0, y'_0, \ldots, y_0^{m-1}) \end{array}\right\}$$

donde hemos introducido las variables auxiliares y_i, que representan funciones con valores en $\mathbb{R}^n$. Todo este sistema se puede expresar como una única ecuación vectorial en las condiciones de la sección anterior. ■

Ejemplo Vamos a calcular todas las soluciones de la ecuación

$$y''(t) = \frac{k}{t}\, y'(t), \quad k \in \mathbb{R},\ t > 0.$$

Obviamente las funciones constantes son soluciones de la ecuación. Si y no es constante existe un punto $t_0 > 0$ tal que $y'(t_0) \neq 0$. Llamemos $y_1 = y'(t)$. En un entorno de t_0 tenemos

$$\frac{y'_1(t)}{y_1(t)} = \frac{k}{t},$$

luego integrando entre t_0 y t queda

$$\log y_1(t) - \log y(t_0) = \log x^k,$$

de donde $y_1(t) = y_1(t_0)x^k$, es decir, $y'(t) = y'(t_0)x^k$. Integrando de nuevo concluimos que

$$y(t) = \begin{cases} \dfrac{y'(t_0)}{k+1}\, x^{k+1} + y(t_0) & \text{si } k \neq -1 \\ y'(t_0) \log t + y(t_0) & \text{si } k = -1 \end{cases}$$

Ahora es claro que las soluciones de la ecuación dada están todas definidas en $]0, +\infty[$ y vienen dadas por

$$y(t) = \begin{cases} Ax^{k+1} + B & \text{si } k \neq -1 \\ A \log t + B & \text{si } k = -1 \end{cases}$$

Estas expresiones incluyen las funciones constantes, que habíamos dejado aparte.

■

El análogo de segundo orden a un campo de velocidades sería un campo de aceleraciones, pero en física resulta más natural hablar de *campos de fuerzas.* Es frecuente que la fuerza $F_t(x)$ que actúa sobre un cuerpo en un instante dado dependa únicamente de su posición en el espacio, con lo que tenemos una función $F : I \times D \longrightarrow \mathbb{R}^n$. El campo de fuerzas será *estacionario* si no depende del tiempo. De acuerdo con la segunda ley de Newton, la trayectoria $x(t)$ de un cuerpo de masa m sometido a este campo vendrá determinada por la ecuación diferencial

$$F_t = m\, x''.$$

La solución depende de la posición inicial x_0 y la velocidad inicial v_0.

Ejemplo La *ley de la gravitación universal* de Newton afirma que la fuerza con que se atraen mutuamente dos cuerpos es directamente proporcional a sus masas e inversamente proporcional al cuadrado de la distancia que los separa. Consideremos una región del espacio donde haya un cuerpo S de masa M tan grande que la masa de cualquier otro cuerpo en las proximidades resulte despreciable. Es el caso del Sol, rodeado de planetas de masa insignificante a su lado, o de la Tierra y sus alrededores. En tal caso podemos suponer que la única fuerza que actúa sobre un cuerpo es la provocada por S. En efecto, cuando dejamos caer un objeto nuestro cuerpo lo atrae por gravedad, pero esta atracción es completamente inapreciable frente a la gravitación terrestre. Igualmente, Júpiter atrae gravitatoriamente a la Tierra, pero la fuerza con que lo hace es insignificante frente a la del Sol.

Si tomamos un sistema de referencia con origen en el punto donde se halla el objeto masivo S, la fuerza que experimenta otro cuerpo de masa m situado en un punto x viene dada por

$$F = -\frac{GMm}{\|x\|^3}\, x,$$

donde $G = 6{,}672 \cdot 10^{-11}\mathrm{N} \cdot \mathrm{m}^2/\mathrm{Kg}^2$ es la *constante de gravitación universal.* El hecho de que su valor sea tan pequeño hace que la gravedad no se manifieste salvo en presencia de grandes masas, como las estrellas y los planetas.

Tras estudiar minuciosamente una gran cantidad de observaciones astronómicas, en 1610 Kepler publicó su *astronomia nova,* donde concluía que los planetas se mueven siguiendo órbitas elípticas, de modo que el Sol ocupa uno de los focos. Ésta es la primera ley de Kepler. Newton mostró que éste y muchos otros hechos sobre el movimiento de los astros pueden deducirse a partir de las

leyes básicas de la dinámica y de su ley de gravitación. Vamos a comprobar que las trayectorias de los objetos sometidos a un campo de fuerzas como el que estamos considerando son rectas o secciones cónicas.

En primer lugar, es claro que si un cuerpo se encuentra en un punto x_0 en las proximidades del Sol con una velocidad v_0, su trayectoria no saldrá del plano determinado por los vectores x_0 y v_0 (o de la recta que determinan, si son linealmente dependientes). Por ello podemos tomar el sistema de referencia de modo que el eje Z sea perpendicular a este plano, con lo que la trayectoria cumplirá $z = 0$ y podemos trabajar únicamente con las coordenadas (x, y).

Conviene introducir coordenadas polares, $r = (x, y) = (\rho \cos\theta, \rho \operatorname{sen}\theta)$. En general, cuando la trayectoria de un móvil viene dada en coordenadas polares por unas funciones $\rho(t)$, $\theta(t)$, se llama *velocidad angular* a la derivada $\omega = \theta'$. La segunda derivada $\alpha = \omega' = \theta''$ se conoce como *aceleración angular.*

La ecuación de Newton puede expresarse en términos de la *cantidad de movimiento,* definida como $p = mv$, donde $v = r'$ es la velocidad del móvil y m es su masa. Admitiendo que ésta es constante, la segunda ley de Newton afirma que

$$F = \frac{dp}{dt},$$

donde F es la fuerza total que actúa sobre el cuerpo. En particular, la cantidad de movimiento de un cuerpo sobre el que no actúa ninguna fuerza permanece constante.

Existen magnitudes análogas a la cantidad de movimiento y la fuerza en coordenadas polares. Se llama *momento angular* de un móvil a la magnitud $L = r \wedge p = m\, r \wedge v$. Si la trayectoria está contenida en un plano el vector L es perpendicular a él. Veamos su expresión en coordenadas polares. Para ello calculamos:

$$v = \rho'(\cos\theta, \operatorname{sen}\theta) + \rho\omega(-\operatorname{sen}\theta, \cos\theta), \tag{6.1}$$

de donde

$$L = m\rho(\cos\theta, \operatorname{sen}\theta, 0) \wedge \rho\omega(-\operatorname{sen}\theta, \cos\theta, 0) = (0, 0, m\rho^2\omega).$$

Para un movimiento plano podemos abreviar y escribir $L = m\rho^2\omega$. Se define el *momento* de una fuerza F que actúa sobre un móvil de posición r como $M = r \wedge F$. Es claro que si sobre un cuerpo actúan varias fuerzas el momento de la fuerza resultante es la suma de los momentos. La versión angular de la segunda ley de Newton es

$$\frac{dL}{dt} = mv \wedge v + r \wedge ma = r \wedge F = M,$$

es decir, el momento total que actúa sobre un móvil es la derivada de su momento angular. En particular, si un cuerpo está libre de toda fuerza su momento angular permanece constante. Sin embargo, el momento angular se conserva incluso en presencia de fuerzas, con tal de que la fuerza resultante sea paralela a la posición, como ocurre en el caso de la fuerza gravitatoria que el Sol ejerce sobre los planetas.

Esto ya nos da una información sobre el movimiento de los planetas: puesto que $m\rho^2\omega$ ha de ser constante, los planetas giran más rápidamente cuando están más cerca del Sol.

Veamos ahora la expresión de la segunda ley de Newton para la gravitación en coordenadas polares. Calculamos la aceleración

$$a = v' = (\rho'' - \rho\omega^2)(\cos\theta, \operatorname{sen}\theta) + (2\rho'\omega + \rho\alpha)(-\operatorname{sen}\theta, \cos\theta).$$

La fuerza gravitatoria es

$$F = -\frac{GMm}{\rho^2}(\cos\theta, \operatorname{sen}\theta).$$

Teniendo en cuenta que los vectores $(\cos\theta, \operatorname{sen}\theta)$ y $(-\operatorname{sen}\theta, \cos\theta)$ son ortogonales,la ecuación $F = ma$ equivale a las ecuaciones

$$\rho'' - \rho\omega^2 = -\frac{GM}{\rho^2}, \quad 2\rho'\omega + \rho\alpha = 0. \tag{6.2}$$

La solución es complicada, pero ahora estamos interesados únicamente en la forma de la trayectoria, aunque sea con otra parametrización. Por ello, en lugar de calcular las funciones $\rho(t)$ y $\theta(t)$ calcularemos la parametrización $\rho(\theta)$. Por supuesto hay un caso en el que esta parametrización es imposible. Si $\omega_0 = 0$ (lo cual equivale a que la velocidad inicial v_0 sea nula o paralela a r_0) entonces es fácil ver que la solución de las ecuaciones es una recta de la forma $\big(\rho(t), \theta_0\big)$, donde ρ está determinada por la ecuación

$$\rho'' = -\frac{GM}{\rho^2}$$

con las condiciones iniciales ρ_0 y ρ_0', determinadas a su vez por r_0 y v_0. En efecto, una trayectoria de este tipo cumple $\omega = \alpha = 0$ y satisface trivialmente las ecuaciones. En definitiva, el cuerpo se aleja del Sol en línea recta o bien cae sobre él. Hemos probado que si la trayectoria de un móvil cumple $\omega(t) = 0$ en un instante t entonces $\omega = 0$ en todo instante, luego, recíprocamente, si $\omega_0 \neq 0$ entonces ω no se anula nunca. Esto hace que la función $\theta(t)$ sea un cambio de parámetro, luego podemos considerar la reparametrización $\rho(\theta)$. Claramente

$$\rho' = \rho_\theta'\omega \quad \text{y} \quad \rho'' = \rho_\theta''\omega^2 + \rho_\theta'\alpha.$$

Sustituimos estas igualdades en (6.2) y eliminamos α en la primera usando la segunda. El resultado es la ecuación

$$\left(\rho'' - \frac{2\rho'^2}{\rho} - \rho\right)\omega^2 = -\frac{GM}{\rho^2},$$

donde ahora todas las derivadas son respecto de θ y no respecto de t. No obstante, la presencia de ω nos obliga a considerar ambos miembros como funciones

de t (la expresión entre paréntesis es una función de θ compuesta con la función $\theta(t)$).

Sin embargo, al multiplicar ambos miembros por $m^2\rho^4$ queda

$$\frac{1}{\rho^2}\left(\rho'' - \frac{2\rho'^2}{\rho} - \rho\right) = -\frac{GMm^2}{L^2}, \tag{6.3}$$

donde $L = m\rho^2\omega$ es una constante, luego todo el segundo miembro es constante y la igualdad sigue siendo válida si consideramos el primer miembro como función de θ.

Recordemos que si r es una recta y F un punto exterior, la cónica de directriz r y foco F está formada por los puntos tales que la razón entre las distancias a r y a F es constante (y recibe el nombre de *excentricidad* de la cónica). Toda cónica que no sea una circunferencia es de esta forma. Si suponemos que el foco es el origen y la directriz es la recta vertical $x = p > 0$, entonces la distancia de un punto de coordenadas polares (ρ, θ) a la directriz es $|p - \rho\cos\theta|$, luego la ecuación de la cónica de excentricidad ϵ es

$$\frac{\rho}{p - \rho\cos\theta} = \epsilon.$$

En principio faltaría un valor absoluto. Si $\epsilon \leq 1$ tenemos una elipse o una parábola y podemos suprimirlo, pues la curva queda al mismo lado de la directriz que el foco. Si $\epsilon > 1$ tenemos una hipérbola, y al suprimir el valor absoluto estamos quedándonos con una de sus ramas. Lo hacemos así porque los cuerpos que se mueven siguiendo trayectorias hiperbólicas tienen al Sol en el foco correspondiente a la rama que siguen o, dicho de otro modo, que la rama que eliminamos no va a ser solución de la ecuación diferencial. Despejando ρ y llamando $r = p\epsilon$ queda

$$\rho = \frac{r}{1 + \epsilon\cos(\theta + k)},$$

La constante k equivale a girar la cónica, de modo que ahora la directriz es arbitraria. Cualquier curva de esta forma es una cónica de excentricidad ϵ. Además esta expresión incorpora también a las circunferencias, para las que $\epsilon = 0$. Es fácil calcular

$$\rho' = \frac{\epsilon\rho^2}{r}\operatorname{sen}(\theta + k), \quad \rho'' = \frac{2\epsilon^2\rho^3}{r^2}\operatorname{sen}^2(\theta + k) + \frac{\epsilon\rho^2}{r}\cos(\theta + k).$$

Al sustituir en el miembro izquierdo de (6.3) se obtiene sin dificultad el valor $-1/r$, luego concluimos que la cónica

$$\rho = \frac{L^2}{GMm^2}\big(1 + \epsilon\cos(\theta + k)\big)^{-1}$$

satisface (6.3). Sólo queda probar que éstas son las únicas soluciones posibles o, lo que es equivalente, que hay una solución de esta forma cualesquiera que sean las condiciones iniciales ρ_0, θ_0, ρ_0', ω_0. Basta ver que las ecuaciones

$$\rho_0 = \frac{L^2}{GMm^2}\big(1 + \epsilon\cos(\theta_0 + k)\big)^{-1}, \quad \rho_0' = \frac{GMm^2\rho_0^2}{L^2}\epsilon\operatorname{sen}(\theta_0 + k)$$

tienen solución en $\epsilon > 0$, k para todos los valores de ρ_0, θ_0, ρ_0', ω_0, pero sustituyendo $L = m\rho_0^2\omega_0$ estas ecuaciones equivalen a

$$\epsilon\big(\cos(\theta_0 + k), \operatorname{sen}(\theta_0 + k)\big) = \left(\frac{\rho_0^3\omega_0^2}{GM} - 1, \frac{\rho_0'\rho_0^2\omega_0^2}{GM}\right),$$

que obviamente tienen solución. Con esto hemos probado que las trayectorias de los objetos sometidos a la atracción de una masa fija puntual son rectas o secciones cónicas. ■

Ejemplo Sea S una variedad diferenciable, sea $p \in S$ y $w \in T_p(S)$ un vector unitario. Sea X una carta alrededor de p, $p = X(x_0)$ y $w = dX(x_0')$. Entonces existe una única curva $x(t)$ que verifica las ecuaciones (5.10) del capítulo anterior con las condiciones iniciales $x(0) = x_0$, $x'(0) = x_0'$. La curva $g(t) = X\big(x(t)\big)$ es una geodésica de S parametrizada por el arco tal que $g(0) = p$ y $g'(0) = w$. Recíprocamente, la representación en la carta de cualquier geodésica que cumpla esto ha de ser solución de las ecuaciones (5.10), luego g es única (salvo cambio de parámetro). En resumen:

> *En una variedad, por cada punto pasa una única geodésica en cada dirección.*

Por ejemplo, en el capítulo anterior probamos que los círculos máximos son geodésicas de la esfera. Puesto que por cada punto y en cada dirección pasa un círculo máximo, concluimos que los círculos máximos son las únicas geodésicas de la esfera. ■

***Ejemplo** En el capítulo anterior demostramos que las rectas elípticas e hiperbólicas son geodésicas de los planos elíptico e hiperbólico. Puesto que por cada punto y en cada dirección pasa una única recta, ahora podemos concluir que las rectas son las únicas geodésicas. ■

Capítulo VII

Teoría de la medida

Pensemos en el concepto de área de una figura plana F. Una primera aproximación consiste en definir dicha área $\mu(F)$ como el número de veces que F contiene al cuadrado de lado unidad, pero esta definición sólo es aplicable si F es unión de un número finito de cuadrados unitarios. Por ejemplo, si F es un rectángulo de lados m y n (naturales) entonces $\mu(F) = mn$. Si F es un rectángulo de lados racionales $r = p/q$ y $r' = p'/q'$, es fácil concluir que el área del rectángulo de lados p y q ha de ser qq' veces el área de F, de donde $\mu(F) = rr'$. Si F es un rectángulo de lados arbitrarios α y β, ya no podemos compararlo directamente con cuadrados unitarios y hemos de recurrir a un argumento de continuidad: si $r < \alpha < r'$, $s < \beta < s'$ son números racionales, el área de F ha de ser mayor que el área de un rectángulo de lados r y s y menor que la de un rectángulo de lados r' y s', es decir, $rs \leq \mu(F) \leq r's'$. El único número real posible es $\mu(F) = \alpha\beta$. Una vez determinada el área de un rectángulo arbitrario, existen numerosos y elegantes argumentos que nos permiten calcular el área de muchas figuras, basados en que el área se conserva por isometrías así como en un principio elemental:

$$\text{Si } F \cap G = \varnothing, \text{ entonces } \mu(F \cup G) = \mu(F) + \mu(G). \tag{7.1}$$

Es fácil calcular el área de un paralelogramo, de un triángulo, de un polígono regular arbitrario, etc. Sin embargo, para calcular el área de figuras más complicadas, como pueda ser un círculo, necesitamos de nuevo argumentos de continuidad. Por ejemplo, si llamamos P_n al polígono regular de 2^n lados inscrito en un círculo C y con un vértice en un punto prefijado, es fácil ver que

$$P_2 \subset P_3 \subset P_4 \subset \cdots C, \quad \text{y} \quad C = \bigcup_{n=2}^{\infty} P_n,$$

y en estas circunstancias es natural considerar que

$$\mu(C) = \sup_n \mu(P_n).$$

Admitiendo (7.1), es fácil ver que el hecho de que el área de la unión de una sucesión creciente de conjuntos sea el supremo de las áreas es equivalente a que el área de una unión disjunta numerable de conjuntos sea la suma de sus áreas. Este principio engloba a (7.1) y es suficiente para justificar todos los razonamientos sobre cálculo de áreas. Con él como única base podemos justificar la existencia y el cálculo de áreas de una amplia familia de figuras. Sin embargo no nos capacita para definir el área de cualquier subconjunto de $\mathbb{R}^2$. Existen problemas técnicos para ello en los que no vamos a entrar, pero lo cierto es que sólo podremos definir una función área $\mu : \mathcal{M} \longrightarrow [0, +\infty]$ sobre una cierta familia $\mathcal{M}$, que contendrá a todos los círculos, triángulos, etc., pero que no será todo el conjunto de partes de $\mathbb{R}^2$.

Conviene introducir una nueva estructura matemática que recoja estas ideas y nos permita extenderlas a otras situaciones análogas (el volumen en $\mathbb{R}^3$, el área en una superficie, etc.)

7.1 Medidas positivas

Definición 7.1 Sea X un conjunto. Un *álgebra* de subconjuntos de X es una familia $\mathcal{A}$ de subconjuntos de X tal que:

a) $\varnothing, X \in \mathcal{A}$.

b) Si $A \in \mathcal{A}$, entonces $X \setminus A \in \mathcal{A}$.

c) Si $A, B \in \mathcal{A}$, entonces $A \cup B, A \cap B \in \mathcal{A}$.

Observar que la propiedad b) hace que la propiedad c) para uniones implique la parte para intersecciones y viceversa. Si dicha propiedad se cumple para familias numerables entonces se dice que $\mathcal{A}$ es una *σ-álgebra.*

Una *medida positiva* (o simplemente una medida) en una σ-álgebra $\mathcal{A}$ de subconjuntos de X es una aplicación $\mu : \mathcal{A} \longrightarrow [0, +\infty]$ que cumpla las propiedades siguientes:

a) $\mu(\varnothing) = 0$.

b) Si $\{A_n\}_{n=0}^{\infty}$ es una familia de conjuntos de $\mathcal{A}$ disjuntos dos a dos, entonces

$$\mu\left(\bigcup_{n=0}^{\infty} A_n\right) = \sum_{n=0}^{\infty} \mu(A_n).$$

Los conjuntos de $\mathcal{A}$ se llaman subconjuntos *medibles* de X. Un *espacio medida* es una terna $(X, \mathcal{A}, \mu)$, donde $\mathcal{A}$ es una σ-álgebra de subconjuntos de X y μ es una medida en $\mathcal{A}$. En la práctica escribiremos X en lugar de $(X, \mathcal{A}, \mu)$.

La medida μ es *unitaria* si $\mu(X) = 1$, es *finita* si $0 < \mu(X) < +\infty$ y es *σ-finita* si $\mu(X) > 0$ y existen conjuntos medibles $\{A_n\}_{n=0}^{\infty}$ de medida finita tales que

$$X = \bigcup_{n=0}^{\infty} A_n.$$

Éste último es el caso del área en el plano o el volumen en el espacio. El área del plano es infinita, pero podemos descomponerlo en una unión numerable de bolas de área finita. También se habla de espacios medida unitarios, finitos o σ-finitos, según sea la medida definida en ellos.

La σ-álgebra más simple es la formada por todos los subconjuntos de X, pero ya hemos comentado que no podremos definir medidas interesantes sobre ella. Es claro que la intersección de una familia de σ-álgebras sobre un conjunto X es de nuevo una σ-álgebra, por lo que dado $G \subset X$ existe una mínima σ-álgebra que contiene a G. Se la llama σ-álgebra generada por G. Si X es un espacio topológico, la σ-álgebra generada por los conjuntos abiertos recibe el nombre de *σ-álgebra de Borel.* Una medida definida sobre la σ-álgebra de Borel de un espacio topológico X recibe el nombre de *medida de Borel* en X.

Las propiedades siguientes de las medidas se deducen inmediatamente de la definición. Usamos el convenio de que si $a \in \mathbb{R}$ entonces $a+\infty = +\infty+\infty = +\infty$.

Teorema 7.2 *Sea X un espacio medida.*

a) Si $A \subset B$ son medibles entonces $\mu(A) \leq \mu(B)$.

b) Si $A \subset B$ son medibles y $\mu(A) < +\infty$, entonces $\mu(B \setminus A) = \mu(B) - \mu(A)$.

c) Si A y B son conjuntos medibles disjuntos $\mu(A \cup B) = \mu(A) + \mu(B)$.

d) Si A y B son medibles entonces $\mu(A \cup B) \leq \mu(A) + \mu(B)$.

e) Si $\{A_n\}_{n=0}^{\infty}$ son medibles entonces

$$\mu\left(\bigcup_{n=0}^{\infty} A_n\right) \leq \sum_{n=0}^{\infty} \mu(A_n).$$

f) Si $\{A_n\}_{n=0}^{\infty}$ son medibles y cada $A_n \subset A_{n+1}$, entonces

$$\mu\left(\bigcup_{n=0}^{\infty} A_n\right) = \sup_n \mu(A_n).$$

g) Si $\{A_n\}_{n=0}^{\infty}$ son medibles, cada $A_{n+1} \subset A_n$ y $\mu(A_0) < +\infty$, entonces

$$\mu\left(\bigcap_{n=0}^{\infty} A_n\right) = \operatorname{ínf}_n \mu(A_n).$$

Por ejemplo, para probar el último apartado aplicamos el anterior a los conjuntos $A_0 \setminus A_n$.

Los conjuntos de medida cero se llaman conjuntos *nulos.* Si A es un conjunto nulo y $B \subset A$ o bien B no es medible o bien es nulo. Sucede que siempre podemos suponer que es nulo, en el sentido de que la σ-álgebra donde está definida una medida siempre se puede extender para que incluya a todos los subconjuntos de los conjuntos nulos. Antes de probar esto conviene definir una medida *completa* como una medida para la cual todos los subconjuntos de un conjunto nulo son medibles (y por lo tanto nulos).

Teorema 7.3 *Sea X un conjunto $\mu : \mathcal{A} \longrightarrow [0, +\infty]$ una medida definida sobre una σ–álgebra de subconjuntos de X. Sea*

$$\mathcal{B} = \{A \subset X \mid \textit{existen } B, C \in \mathcal{A} \textit{ tales que } B \subset A \subset C \textit{ y } \mu(C \setminus B) = 0\}.$$

Entonces $\mathcal{B}$ es una σ-álgebra de subconjuntos de X que contiene a $\mathcal{A}$ y μ se extiende a una única medida en $\mathcal{B}$, que es completa.

Demostración: Si $A \in \mathcal{A}$ es claro que $A \in \mathcal{B}$. Basta tomar $B = C = A$. Así pues $\mathcal{A} \subset \mathcal{B}$, luego en particular $\varnothing, X \in \mathcal{B}$.

Si $A \in \mathcal{B}$, sean $B, C \in \mathcal{A}$ tales que $B \subset A \subset C$ y $\mu(C \setminus B) = 0$. Entonces $X \setminus C \subset X \setminus A \subset X \setminus B$, y claramente $X \setminus C, X \setminus A \in \mathcal{A}$ y

$$\mu\big((X \setminus B) \setminus (X \setminus C)\big) = \mu(C \setminus B) = 0.$$

Por lo tanto $X \setminus A \in \mathcal{B}$. Para probar que $\mathcal{B}$ es una σ-álgebra basta probar que la unión numerable de elementos de $\mathcal{B}$ está en $\mathcal{B}$. Sea, pues, $\{A_n\}_{n=0}^{\infty}$ una familia de elementos de $\mathcal{B}$. Sean $\{B_n\}_n$ y $\{C_n\}_n$ según la definición de $\mathcal{B}$. Entonces

$$\bigcup_{n=0}^{\infty} B_n \subset \bigcup_{n=0}^{\infty} A_n \subset \bigcup_{n=0}^{\infty} C_n,$$

los conjuntos de los extremos están en $\mathcal{A}$ y

$$0 \leq \mu\left(\left(\bigcup_{n=0}^{\infty} C_n\right) \setminus \left(\bigcup_{n=0}^{\infty} B_n\right)\right) \leq \mu\left(\bigcup_{n=0}^{\infty} C_n \setminus B_n\right) = 0.$$

Por lo tanto $\bigcup\limits_{n=0}^{\infty} A_n \in \mathcal{B}$. Esto prueba que $\mathcal{B}$ es una σ-álgebra.

Si $A \in \mathcal{B}$ y B, C son los conjuntos dados por la definición, es claro que $\mu(B) = \mu(C)$. Veamos que podemos extender la medida μ definiendo $\mu(A) = \mu(B) = \mu(C)$. Es claro que esta es la única extensión posible. En efecto, si B' y C' también cumplen la definición, entonces $B \setminus B' \subset C' \setminus B'$, luego $\mu(B \setminus B') \leq \mu(C' \setminus B') = 0$, de donde $\mu(B \setminus B') = 0$ y $\mu(B) = \mu(B')$, luego B y B' dan lugar al mismo valor de $\mu(A)$.

Veamos que la extensión de μ que acabamos de definir es realmente una medida. Para ello tomamos una familia $\{A_n\}_{n=0}^{\infty}$ de elementos de $\mathcal{B}$ disjuntos dos a dos. Sean $\{B_n\}$ y $\{C_n\}$ elementos de $\mathcal{A}$ que satisfagan la definición de $\mathcal{B}$. Según hemos visto, los conjuntos $\bigcup\limits_{n=0}^{\infty} B_n$ y $\bigcup\limits_{n=0}^{\infty} C_n$ justifican que $\bigcup\limits_{n=0}^{\infty} A_n \in \mathcal{B}$ y es claro que los B_n son disjuntos dos a dos, luego

$$\mu\left(\bigcup_{n=0}^{\infty} A_n\right) = \mu\left(\bigcup_{n=0}^{\infty} B_n\right) = \sum_{n=0}^{\infty} \mu(B_n) = \sum_{n=0}^{\infty} \mu(A_n).$$

Es claro que la medida en $\mathcal{B}$ es completa, pues si $A \in \mathcal{B}$ es nulo y $D \subset A$, entonces tomamos $B, C \in \mathcal{A}$ tales que $B \subset A \subset C$ con $\mu(B) = \mu(C) = 0$. Claramente $\varnothing \subset D \subset C$ y $\mu(C \setminus \varnothing) = 0$, luego $D \in \mathcal{B}$. ■

La extensión construida en el teorema anterior se conoce como *compleción* de μ. En vista de esto, normalmente podremos suponer sin pérdida de generalidad que trabajamos con medidas completas. Otra propiedad importante que puede poseer una medida sobre un espacio topológico es la regularidad, que definimos a continuación.

Definición 7.4 Diremos que una medida μ en un espacio topológico X es *regular* si todos los abiertos son medibles, los subespacios compactos tienen medida finita y para todo conjunto medible E se cumple

$$\mu(E) = \text{ínf}\{\mu(V) \mid E \subset V, \quad V \text{ abierto}\}$$

y

$$\mu(E) = \sup\{\mu(K) \mid K \subset E, \quad K \text{ compacto}\}.$$

Diremos que la medida es *casi regular* si la segunda propiedad se cumple al menos cuando $\mu(E) < +\infty$ y cuando E es abierto.

En definitiva una medida es regular si la medida de todo conjunto medible puede aproximarse por la medida de un abierto mayor y de un compacto menor. El concepto de medida casi regular lo introducimos por cuestiones técnicas, pero a continuación probamos que en todos los espacios que nos van a interesar es equivalente a la regularidad.

Diremos que un espacio topológico es *σ-compacto* si es unión numerable de conjuntos compactos. Por ejemplo, todo abierto Ω en $\mathbb{R}^n$ es σ-compacto, pues puede expresarse como unión de los compactos

$$\Omega_k = \{x \in \Omega \mid \|x\| \leq k, \ d(x, \mathbb{R}^n \setminus \Omega) \geq 1/k\}, \quad k = 1, 2, 3, \ldots$$

Teorema 7.5 *Toda medida casi regular en un un espacio σ-compacto es regular.*

Demostración: Supongamos que X es la unión de los compactos $\{K_n\}_{n=1}^{\infty}$. Sustituyendo cada K_n por su unión con los precedentes podemos suponer que si $m \leq n$ entonces $K_m \subset K_n$. Dado un conjunto de Borel B tal que $\mu(B) = +\infty$, tenemos que

$$B = \bigcup_{n=1}^{\infty} B \cap K_n,$$

y como la unión es creciente $\sup\limits_n \mu(B \cap K_n) = \mu(B) = +\infty$. Dado $R > 0$ existe un n tal que $\mu(B \cap K_n) > R + 1$ y, como μ es casi regular, $B \cap K_n$ tiene medida finita y existe un compacto $K \subset B \cap K_n$ tal que $\mu(K) > R$, lo que prueba que μ es regular. ■

Ejercicio: Probar que la compleción de una medida regular es una medida regular.

7.2 Funciones medibles

Si X es un espacio medida, a menudo tendremos que trabajar con subconjuntos de X definidos a partir de una aplicación $f : X \longrightarrow Y$ y necesitaremos garantizar que dichos conjuntos son medibles. Esto lo lograremos mediante el concepto de aplicación medible.

Definición 7.6 Si X es un espacio medida e Y es un espacio topológico, una aplicación $f : X \longrightarrow Y$ es *medible* si las antiimágenes por f de los abiertos de Y son conjuntos medibles.

Como las antiimágenes conservan las operaciones conjuntistas es muy fácil probar que los conjuntos de Y cuyas antiimágenes son medibles forman una σ-álgebra, que en el caso de una función medible contiene a los abiertos, luego contendrá a todos los conjuntos de Borel, es decir, una aplicación es medible si y sólo si las antiimágenes de los conjuntos de Borel son conjuntos medibles. El siguiente caso particular nos interesará especialmente:

Teorema 7.7 *Una aplicación $f : X \longrightarrow [-\infty, +\infty]$ es medible si y sólo si lo son todos los conjuntos $f^{-1}\big[]x, +\infty]\big]$, para todo $x \in \mathbb{R}$.*

Demostración: Ya hemos comentado que los conjuntos con antiimagen medible forman una σ-álgebra $\mathcal{A}$. Hemos de ver que $\mathcal{A}$ contiene a los abiertos de $[-\infty, +\infty]$. Por hipótesis contiene a los intervalos $]x, +\infty]$, luego también a sus complementarios $[-\infty, x]$. Todo intervalo $[-\infty, x[$ es intersección numerable de los intervalos $[-\infty, x + 1/n]$, luego también está en $\mathcal{A}$. De aquí se sigue que $\mathcal{A}$ contiene también a los intervalos $]x, y[\, = \,]x, +\infty] \cap [-\infty, y[$. Finalmente, todo abierto de $[-\infty, +\infty]$ se expresa como unión numerable de intervalos abiertos, luego está en $\mathcal{A}$. ■

Es claro que la composición de una función medible con una función continua es una función medible. Esto nos da, por ejemplo, que si $f : X \longrightarrow [-\infty, +\infty]$ es medible, también lo es $|f|$ y αf para todo número real α, así como $1/f$ si f no se anula.

Para probar resultados análogos cuando intervienen dos funciones (suma de funciones medibles, etc.) usaremos la observación siguiente:

> *Si los espacios topológicos Y, Z tienen bases numerables (como $[-\infty, +\infty]$ y sus subespacios) y $u : X \longrightarrow Y$, $v : X \longrightarrow Z$ son aplicaciones medibles, entonces la aplicación $u \times v : X \longrightarrow Y \times Z$ dada por $(u \times v)(x) = \big(u(x), v(x)\big)$ es medible.*

Basta observar que los productos de abiertos básicos $A \times B$ forman una base numerable de $Y \times Z$, luego todo abierto de $Y \times Z$ es unión numerable de estos conjuntos, por lo que es suficiente que sus antiimágenes sean medibles, pero $(u \times v)^{-1}[A \times B] = u^{-1}[B] \cap v^{-1}[C]$.

Ahora, por ejemplo, si $u, v : X \longrightarrow \mathbb{R}$ son aplicaciones medibles, también lo son $f + g$ y fg, pues son la composición de $u \times v$ con la suma y el producto, que son continuas.

Nos interesa extender este resultado a funciones $u, v : X \longrightarrow [-\infty, +\infty]$, pero entonces tenemos el problema de que no es posible extender la suma y el producto de modo que sean continuas en los puntos $(+\infty, -\infty)$, $(-\infty, +\infty)$ en el caso de la suma y en los puntos $(\pm\infty, 0)$, $(0, \pm\infty)$ en el caso del producto.

Hacemos esto: definimos $u + v$ de modo que $+\infty - \infty = 0$ (por ejemplo) y ahora observamos lo siguiente:

> *Sea $u : X \longrightarrow Y$ una función medible, A un subconjunto medible de X e $y \in Y$. Entonces la función $v : X \longrightarrow Y$ que coincide con u fuera de A y toma el valor y en A es medible.*

La razón es que

$$v^{-1}[B] = \begin{cases} u^{-1}[B] \setminus A & \text{si } y \notin B \\ u^{-1}[B] \cup A & \text{si } y \in B \end{cases}$$

Así, dadas $u, v : X \longrightarrow [-\infty, +\infty]$ medibles tales que donde una vale $+\infty$ la otra no vale $-\infty$, las modificamos para que valgan 0 donde toman valores infinitos, las sumamos y obtenemos una función medible, luego modificamos la suma para que tome el valor ∞ adecuado donde deba tomar dichos valores (claramente en un conjunto medible), con lo que obtenemos una función medible. Igualmente con el producto.

Otra consecuencia del teorema sobre el producto cartesiano de funciones medibles es que si tenemos dos funciones medibles $u, v : X \longrightarrow [-\infty, +\infty]$, los conjuntos del estilo de $\{x \in X \mid u(x) < v(x)\}$ son medibles (por ejemplo en este caso se trata de la antiimagen por $u \times v$ del abierto $\{(x, y) \mid x < y\}$).

Dos operaciones definibles sobre todas las funciones $f, g : X \longrightarrow [-\infty, +\infty]$ son las dadas por $(f \vee g)(x) = \text{máx}\{f(x), g(x)\}$ y $(f \wedge g)(x) = \text{mín}\{f(x), g(x)\}$.

Las funciones $\vee, \wedge : [-\infty, +\infty] \times [-\infty, +\infty] \longrightarrow [-\infty, +\infty]$ son ambas continuas, pues lo son trivialmente cuando se restringen a los cerrados determinados por las condiciones $x \leq y$ e $y \leq x$, respectivamente. Esto implica que si $f, g : X \longrightarrow [-\infty, +\infty]$ son medibles, también lo son $f \vee g$ y $f \wedge g$.

En particular si $f : X \longrightarrow [-\infty, +\infty]$ es una función medible, definimos las funciones $f^+ = f \vee 0$ y $f^- = -(f \wedge 0)$, llamadas *parte positiva* y *parte negativa* de f, respectivamente.

Tenemos que si f es medible también lo son f^+ y f^-. El recíproco es cierto porque claramente $f = f^+ - f^-$. Además $|f| = f^+ + f^-$.

Seguidamente probaremos que la medibilidad se conserva al tomar límites. Si $\{a_n\}_{n=0}^{\infty}$ es una sucesión en $[-\infty, +\infty]$, definimos sus límites superior e inferior como

$$\overline{\lim_n}\, a_n = \inf_{k \geq 0} \sup_{n \geq k} a_n, \quad \underline{\lim_n}\, a_n = \sup_{k \geq 0} \inf_{n \geq k} a_n.$$

En el capítulo III demostrábamos que el límite superior de una sucesión es el supremo de sus puntos adherentes. Igualmente se prueba que el límite inferior es el ínfimo de sus puntos adherentes.

Una sucesión converge si y sólo si tiene un único punto adherente (su límite), por lo que $\{a_n\}_{n=0}^\infty$ converge si y sólo si $\overline{\lim_n} a_n = \underline{\lim_n} a_n$ y entonces

$$\overline{\lim_n} a_n = \underline{\lim_n} a_n = \lim_n a_n.$$

Si $\{f_n\}_{n=0}^\infty$ es una sucesión de funciones $f_n : X \longrightarrow [-\infty, +\infty]$, definimos puntualmente las funciones $\sup_n f_n$, $\inf_n f_n$, $\underline{\lim_n} f_n$ y $\overline{\lim_n} f_n$. Si la sucesión es puntualmente convergente las dos últimas funciones coinciden con la función límite puntual $\lim_n f_n$.

Teorema 7.8 *Si las funciones f_n son medibles, también lo son las funciones $\sup_n f_n$, $\inf_n f_n$, $\overline{\lim_n} f_n$ y $\underline{\lim_n} f_n$.*

Demostración: Claramente

$$\left(\sup_n f_n\right)^{-1} \big[]x, +\infty]\big] = \bigcup_{n=0}^\infty f_n^{-1}\big[]x, +\infty]\big],$$

es medible. Igualmente se prueba con ínfimos y de aquí se deducen los resultados sobre límites superiores e inferiores. En particular, el límite puntual de una sucesión de funciones medibles es una función medible. ■

Si X es un espacio medida y E es un subconjunto medible, entonces los subconjuntos medibles de E forman una σ-álgebra de subconjuntos de E y la medida de X restringida a esta σ-álgebra es una medida en E. En lo sucesivo consideraremos a todos los subconjuntos medibles de los espacios medida como espacios medida de esta manera.

Notar que si $X = \bigcup_{n=0}^\infty E_n$ es una descomposición de X en subconjuntos medibles (no necesariamente disjuntos), entonces $f : X \longrightarrow Y$ es medible si y sólo si lo son todas las funciones $f|_{E_n}$, pues si f es medible y G es un abierto en Y, $(f|_{E_n})^{-1}[G] = f^{-1}[G] \cap E_n$, luego $(f|_{E_n})^{-1}[G]$ es medible, y si las $f|_{E_n}$ son medibles, entonces $f^{-1}[G] = (f|_{E_n})^{-1}[G]$, luego también es medible.

Si E es un subconjunto de X, su función característica χ_E es medible si y sólo si lo es E.

Si extendemos una función medible $f : E \longrightarrow [-\infty, +\infty]$ asignándole el valor 0 fuera de E, obtenemos una función medible en X. Por ello identificaremos las funciones medibles $f : E \longrightarrow [-\infty, +\infty]$ con las funciones medibles en X que se anulan fuera de E. En particular identificaremos la restricción a E de una función $f : X \longrightarrow [-\infty, +\infty]$ con la función $f\chi_E$.

Los resultados que hemos dado son suficientes para garantizar que todas las funciones que manejaremos y los conjuntos definidos por ellas son medibles. No insistiremos en ello a menos que haya alguna dificultad inusual.

7.3 La integral de Lebesgue

En todo espacio medida X podemos definir una integral que generaliza fuertemente a la integral de Riemann que estudiamos en el capítulo anterior para el caso de la recta real. Más adelante veremos que es posible definir una única medida en $\mathbb{R}$ de modo que $\mu\big([a,b]\big) = b-a$. La integral asociada a esta medida coincide con la integral de Riemann sobre las funciones en las que ésta está definida.

La idea fundamental es que el papel que juegan los intervalos en la integral de Riemann lo juegan los conjuntos medibles en el caso general. Al considerar conjuntos más generales obtenemos muchas más funciones integrables, lo que se traduce en que la integrabilidad se conserva no sólo por las operaciones algebraicas (como en el caso de la integral de Riemann) sino también por pasos al límite. En lugar de dar una definición de integral que muestre estas ideas, aprovecharemos las propiedades de convergencia para dar una definición rápida.

Definición 7.9 Una función *simple* en un espacio medida X es una función medible $s : X \longrightarrow [0,+\infty[$ que sólo toma un número finito de valores $\alpha_1,\ldots,\alpha_n$. Si llamamos $A_i = s^{-1}[\alpha_i]$, entonces los conjuntos A_i son medibles disjuntos y $s = \sum\limits_{i=1}^{n} \alpha_i \chi_{A_i}$.

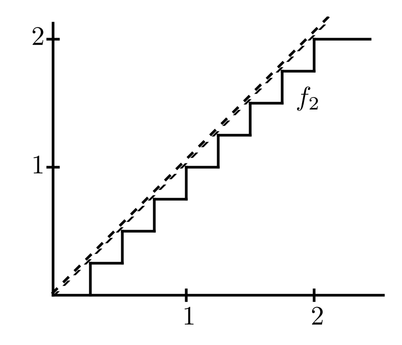
La base de nuestra construcción de la integral será el teorema siguiente:

Teorema 7.10 *Si X es un espacio medida y $f : X \longrightarrow [0,+\infty]$ es una función medible, entonces existe una sucesión $\{s_n\}_{n=1}^{\infty}$ de funciones simples en X tal que*

$$0 \leq s_1 \leq s_2 \leq \cdots \leq f \quad \text{y} \quad f = \lim_n s_n.$$

DEMOSTRACIÓN: Para cada número natural $n > 0$ y cada $t \in \mathbb{R}$ existe un único $k = k_n(t) \in \mathbb{N}$ tal que $k/2^n \leq t < (k+1)/2^n$. Sea $f_n : [0,+\infty] \longrightarrow [0,+\infty]$ dada por

$$f_n(t) = \begin{cases} k_n(t)/2^n & \text{si } 0 \leq t < n \\ n & \text{si } n \leq t \leq +\infty \end{cases}$$



La figura muestra la función f_2. Claramente f_n toma un número finito de valores y

$$f_1 \leq f_2 \leq f_3 \leq \cdots \leq I,$$

donde I es la función identidad $I(t) = t$. Como $t - 1/2^n < f_n(t) \leq t$ para $0 \leq t \leq n$, es claro que $\{f_n\}_{n=1}^{\infty}$ converge puntualmente a I.

Sea $s_n = f \circ f_n$. Claramente s_n toma un número finito de valores (a lo sumo los que toma f_n) y las antiimágenes de estos valores son las antiimágenes por f de los intervalos donde los toma f_n, luego son conjuntos medibles. Además

$$0 \leq s_1 \leq s_2 \leq \cdots \leq f,$$

luego las funciones s_n son simples y, tomando límites, es obvio que la sucesión converge puntualmente a f. ■

Definición 7.11 Sea X un espacio medida y $s = \sum\limits_{i=1}^n \alpha_i \chi_{A_i}$ una función simple en X. Definimos la *integral* de s en X como

$$\int_X s\, d\mu = \sum_{i=1}^n \alpha_i \mu(A_i) \in [0, +\infty],$$

con el convenio de que $+\infty \cdot 0 = 0$.

Si E es un subconjunto medible de X, entonces $s|_E = \sum\limits_{i=1}^n \alpha_i \chi_{A_i \cap E}$, donde las funciones características se toman ahora sobre E. Por lo tanto

$$\int_E s|_E\, d\mu = \sum_{i=1}^n a_i\, \mu(A_i \cap E).$$

Por otro lado $s\chi_E = \sum\limits_{i=1}^n \alpha_i \chi_{A_i \cap E}$ (con las funciones características en X), luego concluimos que

$$\int_E s\, d\mu = \int_X s\chi_E\, d\mu = \sum_{i=1}^n \alpha_i\, \mu(A_i \cap E),$$

es decir, que a efectos de integración podemos adoptar consistentemente el convenio explicado antes por el que identificamos la función $s|_E$ con $s\chi_E$.

Ahora necesitamos el siguiente resultado técnico, que después generalizaremos notablemente.

Teorema 7.12 *Sea X un espacio medida.*

a) Sea s una función simple en X. Para cada subconjunto medible E de X definimos $\nu(E) = \int_E s\, d\mu$. Entonces ν es una medida en X.

b) Si s y t son funciones simples en X se cumple

$$\int_X (s+t)\, d\mu = \int_X s\, d\mu + \int_X t\, d\mu.$$

Demostración: a) Sea $s = \sum\limits_{i=1}^n \alpha_i \chi_{A_i}$. Claramente $\nu(\varnothing) = 0$. Sea $E = \bigcup\limits_{j=1}^\infty E_j$ una unión disjunta de conjuntos medibles. Entonces

$$\begin{aligned} \nu(E) &= \sum_{i=1}^n \alpha_i \mu(A_i \cap E) = \sum_{i=1}^n \alpha_i \sum_{j=1}^\infty \mu(A_i \cap E_j) \\ &= \sum_{j=1}^\infty \sum_{i=1}^n \alpha_i \mu(A_i \cap E_j) = \sum_{j=1}^\infty \nu(E_j). \end{aligned}$$

b) Sean $s = \sum\limits_{i=1}^{n} \alpha_i \chi_{A_i}$ y $t = \sum\limits_{j=1}^{m} \beta_j \chi_{B_j}$. Llamemos $E_{ij} = A_i \cap B_j$. Así, tanto s como t son constantes en los conjuntos E_{ij} (s toma el valor α_i y t el valor β_j). Por lo tanto

$$\int_{E_{ij}} (s+t)\, d\mu = (\alpha_i + \beta_j)\mu(E_{ij}) = \alpha_i \mu(E_{ij}) + \beta_j \mu(E_{ij}) = \int_{E_{ij}} s\, d\mu + \int_{E_{ij}} t\, d\mu.$$

Como los conjuntos E_{ij} son disjuntos dos a dos y su unión es X, la parte a) nos da que la igualdad se cumple para integrales en X. ■

En particular notamos que si $s \leq t$ son funciones simples en un espacio medida X, entonces $t - s$ también es una función simple y

$$\int_X s\, d\mu \leq \int_X s\, d\mu + \int_X (t-s)\, d\mu = \int_X t\, d\mu.$$

En particular se cumple que

$$\int_X t\, d\mu = \sup\Big\{\int_X s\, d\mu \mid s \text{ es una función simple, } s \leq t\Big\}.$$

Esto hace consistente la siguiente definición:

Definición 7.13 Sea X un espacio medida y $f : X \longrightarrow [0, +\infty]$ una función medible. Definimos la *integral* de f como

$$\int_X f\, d\mu = \sup\Big\{\int_X s\, d\mu \mid s \text{ es una función simple, } s \leq f\Big\} \in [0, +\infty].$$

Observar que si E es un subconjunto medible de X, si s es una función simple en E por debajo de $f|_E$, su extensión a X (nula fuera de E) es una función simple bajo $f\chi_E$, y la restricción a E de una función simple en X bajo $f\chi_E$ es una función simple en E bajo $f|_E$. De aquí se sigue que

$$\int_E f\, d\mu = \int_X f\chi_E\, d\mu,$$

pues ambas integrales son el supremo del mismo conjunto de números reales.

Las propiedades siguientes son inmediatas a partir de la definición:

Teorema 7.14 *Sea X un espacio medida y E un subconjunto medible de X.*

a) Si $0 \leq f \leq g$ son funciones medibles en X, entonces $\int_X f\, d\mu \leq \int_X g\, d\mu$.

b) Si $f \geq 0$ es una función medible en X y $A \subset B$ son subconjuntos medibles de X, entonces $\int_A f\, d\mu \leq \int_B f\, d\mu$.

c) Si $f \geq 0$ es una función medible en X y $f|_E = 0$, entonces $\int_E f\, d\mu = 0$ (aunque sea $\mu(E) = +\infty$).

d) Si $f \geq 0$ es una función medible en X y $\mu(E) = 0$, entonces $\int_E f\,d\mu = 0$ (aunque sea $f|_E = +\infty$).

El resultado siguiente es uno de los más importantes del cálculo integral:

Teorema 7.15 (de la convergencia monótona de Lebesgue) *Sea X un espacio medida y $\{f_n\}_{n=1}^\infty$ una sucesión de funciones medibles en X tal que*

$$0 \leq f_1 \leq f_2 \leq \cdots \leq f \quad y \quad f = \lim_n f_n.$$

Entonces f es medible y

$$\int_X f\,d\mu = \lim_n \int_X f_n\,d\mu.$$

Demostración: Por el teorema anterior $\int_X f_n\,d\mu \leq \int_X f_{n+1}\,d\mu$. Toda sucesión monótona creciente en $[0, +\infty]$ converge a su supremo, luego existe $\alpha = \lim\limits_n \int_X f_n\,d\mu \in [0, +\infty]$.

Sabemos que f es medible por ser límite puntual de funciones medibles. De nuevo por el teorema anterior $\int_X f_n\,d\mu \leq \int_X f\,d\mu$, luego $\alpha \leq \int_X f\,d\mu$.

Sea s una función simple $s \leq f$ y sea $0 < c < 1$. Definimos

$$E_n = \{x \in X \mid f_n(x) \geq cs(x)\}, \quad \text{para } n = 1, 2, 3, \ldots$$

Claramente $E_1 \subset E_2 \subset E_3 \subset \cdots$, son conjuntos medibles y, según veremos enseguida, $X = \bigcup\limits_n E_n$.

En efecto, si $x \in X$ y $f(x) = 0$, entonces $x \in E_1$ y si, por el contrario, $f(x) > 0$ entonces $cs(x) < s(x) \leq f(x)$, luego $x \in E_n$ para algún n. Claramente

$$\int_X f_n\,d\mu \geq \int_{E_n} f_n\,d\mu \geq c\int_{E_n} s\,d\mu.$$

Ahora aplicamos el teorema 7.12 y el hecho de que la medida de la unión de una sucesión creciente de conjuntos es el supremo de las medidas, con lo que obtenemos

$$\alpha = \lim_n \int_X f_n\,d\mu \geq c\lim_n \int_{E_n} s\,d\mu = c\int_X s\,d\mu.$$

Como esto es cierto para todo $c < 1$ podemos concluir que $\alpha \geq \int_X s\,d\mu$ para toda función simple $s \leq f$, luego tomando el supremo de estas integrales resulta $\alpha \geq \int_X f\,d\mu$, con lo que tenemos la igualdad buscada. ■

El teorema de la convergencia monótona permite en particular reducir propiedades de la integral de funciones no negativas a propiedades de funciones simples, como ilustra el teorema siguiente, que muestra que la integral conserva las sumas, incluso las infinitas.

Teorema 7.16 *Sea X un espacio medida y sea $\{f_n\}_{n=1}^{\infty}$ una sucesión de funciones no negativas medibles en X. Entonces*

$$\int_X \sum_{n=1}^{\infty} f_n \, d\mu = \sum_{n=1}^{\infty} \int_X f_n \, d\mu.$$

DEMOSTRACIÓN: Probaremos en primer lugar que si f y g son medibles y no negativas, entonces

$$\int_X (f+g)\, d\mu = \int_X f\, d\mu + \int_X g\, d\mu.$$

Tomamos dos sucesiones monótonas $\{s_n\}_{n=1}^{\infty}$ y $\{t_n\}_{n=1}^{\infty}$ de funciones simples convergentes a f y g respectivamente (existen por el teorema 7.10).

Por el teorema 7.12 sabemos que $\int_X (s_n + t_n)\, d\mu = \int_X s_n\, d\mu + \int_X t_n\, d\mu$ y, tomando límites, el teorema de la convergencia monótona nos da la igualdad buscada. En el caso general sabemos, por lo que acabamos de probar, que

$$\int_X \sum_{n=1}^{k} f_n \, d\mu = \sum_{n=1}^{k} \int_X f_n \, d\mu \quad \text{para } k = 1, 2, 3, \ldots$$

Las funciones $\sum\limits_{n=1}^{k} f_n$ forman una sucesión monótona de funciones medibles, luego por el teorema de la convergencia monótona

$$\int_X \sum_{n=1}^{\infty} f_n \, d\mu = \sum_{n=1}^{\infty} \int_X f_n \, d\mu.$$

■

Generalizamos ahora la primera parte del teorema 7.12.

Teorema 7.17 *Sea X un espacio medida y $f : X \longrightarrow [0, +\infty]$ una función medible. Para cada subconjunto medible E de X definimos $\nu(E) = \int_E f\, d\mu$. Entonces ν es una medida en X.*

DEMOSTRACIÓN: Claramente $\nu(\varnothing) = 0$. Sea $E = \bigcup\limits_{n=1}^{\infty} E_n$ una unión disjunta de conjuntos medibles. Es claro que $f\chi_E = \sum\limits_{n=1}^{\infty} f\chi_{E_n}$. Aplicando el teorema anterior queda $\nu(E) = \sum\limits_{n=1}^{\infty} \nu(E_n)$. ■

Después necesitaremos el hecho siguiente:

Teorema 7.18 (Lema de Fatou) *Sea X un espacio medida y sea $\{f_n\}_{n=1}^{\infty}$ una sucesión de funciones medibles no negativas en X. Entonces*

$$\int_X \varliminf_n f_n \, d\mu \leq \varliminf_n \int_X f_n \, d\mu.$$

Demostración: Sea $g_k = \inf_{n\geq k} f_n$. Entonces $g_k \leq f_n$ para $n \geq k$, luego $\int_X g_k\, d\mu \leq \inf_{n\geq k} \int_X f_n\, d\mu$. Además las funciones g_k forman una sucesión monótona creciente que converge a $\underline{\lim}_n f_n$, luego por el teorema de la convergencia monótona

$$\int_X \underline{\lim}_n f_n\, d\mu = \lim_k \int_X g_k\, d\mu = \sup_{k\geq 1} \int_X g_k\, d\mu \leq \sup_{k\geq 1} \inf_{n\geq k} \int_X f_n\, d\mu = \underline{\lim}_n \int_X f_n\, d\mu.$$

■

Ahora extendemos la integral a funciones medibles no necesariamente mayores o iguales que 0.

Definición 7.19 Sea X un espacio medida y $f : X \longrightarrow [-\infty, +\infty]$ una función medible. Entonces f^+ y f^- son funciones medibles no negativas y $f = f^+ - f^-$. Diremos que f es *integrable Lebesgue* en X si tanto $\int_X f^+\, d\mu$ como $\int_X f^-\, d\mu$ son finitas. En tal caso definimos la *integral de Lebesgue* de f como

$$\int_X f\, d\mu = \int_X f^+\, d\mu - \int_X f^-\, d\mu \in \mathbb{R}.$$

Llamaremos $L^1(\mu)$ al conjunto de las funciones integrables Lebesgue en X respecto a la medida μ.

Si una función f es no negativa, entonces $f^- = 0$ y su integral es la que ya teníamos definida. Las propiedades siguientes son todas inmediatas a partir de los resultados que ya hemos demostrado.

Teorema 7.20 *Sea X un espacio medida y sean f, $g : X \longrightarrow [-\infty, +\infty]$ funciones medibles.*

a) f es integrable si y sólo si $\int_X |f|\, d\mu < +\infty$, y en tal caso

$$\left| \int_X f\, d\mu \right| \leq \int_X |f|\, d\mu.$$

b) Si $\alpha, \beta \in \mathbb{R}$, y f, g son integrables, entonces $\alpha f + \beta g$ es integrable y

$$\int_X (\alpha f + \beta g)\, d\mu = \alpha \int_X f\, d\mu + \beta \int_X g\, d\mu.$$

c) Si $f \leq g$ y ambas son integrables, entonces $\int_X f\, d\mu \leq \int_X g\, d\mu$.

d) Si E es un subconjunto medible de X y f es integrable en X, entonces f es integrable en E y $\int_E f\, d\mu = \int_X f\chi_E\, d\mu$.

e) Si E y F son subconjuntos medibles disjuntos de X, entonces la función f es integrable en $E \cup F$ si y sólo si lo es en E y en F y, en tal caso,

$$\int_{E\cup F} f\, d\mu = \int_E f\, d\mu + \int_F f\, d\mu.$$

f) Si E es un subconjunto medible de X y $f|_E = 0$, entonces $\int_E f\,d\mu = 0$.

g) Si E es un subconjunto nulo de X, entonces $\int_E f\,d\mu = 0$.

h) Si f es integrable en X, entonces el conjunto de los puntos donde f toma los valores $\pm\infty$ es nulo.

(La propiedad e sale de aplicar el teorema 7.17 a las partes positiva y negativa de f.)

Algunas consecuencias: por d) vemos que $\int_E 1\,d\mu = \int_X \chi_E\,d\mu = \mu(E)$. Por a) tenemos que si $|f| \leq g$ y g es integrable entonces f también lo es. Toda función medible y acotada sobre un conjunto de medida finita es integrable.

Otra observación de interés es la siguiente: si pasamos de una medida a su compleción, es claro que las funciones simples para la primera lo son también para la segunda y las integrales coinciden. El teorema de la convergencia monótona implica entonces que toda función positiva integrable para una medida sigue siéndolo para su compleción, y de aquí se sigue inmediatamente el resultado para funciones arbitrarias. De este modo la integral respecto a la compleción extiende a la integral respecto a la medida de partida.

Veamos ahora un teorema de convergencia válido para funciones medibles arbitrarias.

Teorema 7.21 (de la convergencia dominada de Lebesgue) *Sea X un espacio medida y sean $\{f_n\}_{n=1}^\infty$ funciones medibles de X en $[-\infty,+\infty]$ que convergen puntualmente a una función f. Si existe una función integrable $g : X \longrightarrow [-\infty,+\infty]$ tal que $|f_n| \leq g$ para todo n, entonces f es integrable y*

$$\int_X f\,d\mu = \lim_n \int_X f_n\,d\mu.$$

Se dice que las funciones f_n están dominadas por g.

Demostración: Claramente $|f| \leq g$, luego f es integrable. Puesto que $|f_n - f| \leq 2g$, podemos aplicar el lema de Fatou a las funciones no negativas $2g - |f_n - f|$, con lo que obtenemos que

$$\int_X 2g\,d\mu \leq \underline{\lim_n} \int_X (2g - |f_n - f|)\,d\mu = \int_X 2g\,d\mu + \underline{\lim_n} \int_X (-|f_n - f|)\,d\mu.$$

Es fácil ver que el signo negativo sale del límite, pero cambiando éste por un límite superior, así $-\overline{\lim_n} \int_X |f_n - f|\,d\mu \geq 0$, o sea, $\overline{\lim_n} \int_X |f_n - f|\,d\mu \leq 0$.

Pero es obvio que $0 \leq \underline{\lim_n} \int_X |f_n - f|\,d\mu \leq \overline{\lim_n} \int_X |f_n - f|\,d\mu = 0$, luego los límites superior e inferior coinciden, luego $\lim_n \int_X |f_n - f|\,d\mu = 0$. Ahora aplicamos que

$$\left| \int_X f_n\,d\mu - \int_X f\,d\mu \right| = \left| \int_X (f_n - f)\,d\mu \right| \leq \int_X |f_n - f|\,d\mu,$$

de donde se sigue el teorema. ■

Cuando una propiedad se verifica para todos los puntos de un espacio medida salvo los de un conjunto nulo se dice que la propiedad se verifica *para casi todo punto,* y lo abreviaremos p.c.t.p. Veamos un ejemplo:

Teorema 7.22 *Si X es un espacio medida y $f : X \longrightarrow [0, +\infty]$ es una función medible tal que $\int_X f\,d\mu = 0$, entonces $f = 0$ p.c.t.p. de X.*

Demostración: Para cada natural $n > 0$ sea $E_n = \{x \in E \mid f(x) > 1/n\}$. Entonces

$$\frac{1}{n}\mu(E_n) \leq \int_{E_n} f\,d\mu \leq \int_X f\,d\mu = 0,$$

luego $\mu(E_n) = 0$. La unión de los E_n es el conjunto $E = \{x \in X \mid f(x) > 0\}$, luego $f = 0$ salvo en los puntos del conjunto nulo E. ■

Cuando digamos que una función $f : X \longrightarrow [-\infty, +\infty]$ está definida p.c.t.p. esto significará que en realidad es $f : X \setminus E \longrightarrow [-\infty, +\infty]$, donde E es un conjunto nulo. Diremos que f es medible si lo es al extenderla a X tomando el valor 0 en E. En tal caso podemos hablar de $\int_X f\,d\mu$ definida como la integral de dicha extensión.

Terminamos la sección con un importante teorema sobre integrales paramétricas:

Teorema 7.23 *Sea U abierto en $\mathbb{R}^n$, K un espacio métrico compacto, μ una medida de Borel finita en K, sean $f : U \times K \longrightarrow \mathbb{R}$ una función continua y $g : K \longrightarrow \mathbb{R}$ una función medible acotada. Definamos $F : U \longrightarrow \mathbb{R}^n$ como la función dada por*

$$F(x) = \int_K f(x, y)g(y)\,d\mu(y),$$

donde $d\mu(y)$ indica que la integral se realiza respecto a la variable $y \in K$, considerando constante a x. Entonces F es continua en U y si existe

$$\frac{\partial f}{\partial x_i} : U \times K \longrightarrow \mathbb{R}$$

y es continua en $U \times K$, entonces existe

$$\frac{\partial F}{\partial x_i} = \int_K \frac{\partial f}{\partial x_i}(x, y)g(y)\,d\mu(y)$$

y es continua en U.

Demostración: Tomemos $x_0 \in U$ y sea B una bola cerrada de centro x_0 contenida en U. Sea M una cota de g en K. Como f es uniformemente continua en $B \times K$, dado $\epsilon > 0$ existe un $\delta > 0$ tal que si $\|x - x_0\| < \delta$, entonces $|f(x, y) - f(x_0, y)| < \epsilon/M\mu(K)$, para todo $y \in K$. Por consiguiente, si $\|x - x_0\| < \delta$ se cumple

$$|F(x) - F(x_0)| \leq \int_K |f(x, y) - f(x_0, y)|\,|g(y)|\,d\mu(y) \leq \epsilon.$$

Esto prueba que F es continua en x_0.

Supongamos ahora la hipótesis de derivabilidad respecto a x_i y sea e_i el i-ésimo vector de la base canónica de $\mathbb{R}^n$. Como $\partial f/\partial x_i$ es uniformemente continua en $B \times K$, existe un $\delta > 0$ tal que si $|h| < \delta$ entonces

$$\left|\frac{\partial f}{\partial x_i}(x_0 + he_i, y) - \frac{\partial f}{\partial x_i}(x_0, y)\right| < \frac{\epsilon}{M\mu(K)}, \quad \text{para todo } y \in K.$$

Si $|h| < \delta$ e $y \in K$, el teorema del valor medio nos da que existe un $r \in \mathbb{R}$ tal que $|r| < |h|$ y

$$f(x_0 + he_i, y) - f(x_0, y) = \frac{\partial f}{\partial x_i}(x_0 + re_i, y)h.$$

(Notar que r depende de y.) Por consiguiente,

$$\left|\frac{f(x_0 + he_i, y)g(y) - f(x_0, y)g(y)}{h} - \frac{\partial f}{\partial x_i}(x_0, y)g(y)\right|$$

$$= \left|\frac{\partial f}{\partial x_i}(x_0 + re_i, y) - \frac{\partial f}{\partial x_i}(x_0, y)\right| |g(y)| < \frac{\epsilon}{\mu(K)}.$$

De aquí se sigue claramente que

$$\left|\frac{F(x_0 + he_i) - F(x_0)}{h} - \int_K \frac{\partial f}{\partial x_i}(x_0, y)g(y)\, d\mu(y)\right| < \epsilon.$$

siempre que $|h| < \delta$, luego existe

$$\frac{\partial F}{\partial x_i}(x_0) = \lim_{h\to 0} \frac{F(x_0 + he_i) - F(x_0)}{h} = \int_K \frac{\partial f}{\partial x_i}(x_0, y)g(y)\, d\mu(y).$$

Además la derivada es continua por la primera parte de este mismo teorema.

■

7.4 El teorema de Riesz

Hasta ahora no tenemos ninguna medida de interés a la que aplicar los resultados que acabamos de exponer. La construcción de medidas es el punto más delicado de toda la teoría. Puesto que el proceso es complicado cualquiera que sea el camino que tomemos, seguiremos uno que nos proporcionará un teorema notable de la teoría de la medida, el teorema de representación de Riesz. Necesitamos unos preliminares topológicos sobre espacios localmente compactos.

Definición 7.24 Un espacio (de Hausdorff) X es *localmente compacto* si todo punto de X tiene una base de entornos compactos.

Así pues, si p es un punto de un espacio localmente compacto X y V es un entorno abierto de p, existe un entorno compacto $K \subset V$. Por definición de entorno existe un abierto W tal que $p \in W \subset K \subset V$ y claramente $\overline{W} \subset K$ es un entorno compacto de p. En resumen, si X es localmente compacto, $p \in X$ y V es un abierto tal que $p \in V$, existe otro abierto W con clausura compacta tal que $p \in W \subset \overline{W} \subset V$.

El teorema siguiente recoge un par de propiedades sencillas:

Teorema 7.25 *Sea X un espacio de Hausdorff.*

a) Si $K \subset X$ es compacto y $p \in X \setminus K$, entonces existen abiertos disjuntos U y V tales que $p \in U$ y $K \subset V$.

b) Si X es localmente compacto, V es abierto en X y $K \subset V$ es compacto, entonces existe un abierto W tal que $K \subset W \subset \overline{W} \subset V$ y $\overline{W}$ es compacto.

DEMOSTRACIÓN: a) Por la propiedad de Hausdorff, para cada $x \in K$ podemos encontrar abiertos disjuntos U_x y V_x tales que $p \in U_x$ y $x \in V_x$. Entonces K está cubierto por los V_x, luego podemos tomar un subcubrimiento finito $V_{x_1}, \ldots, V_{x_n}$. Basta tomar como U la intersección de los U_{x_i} y como V la unión de los V_{x_i}.

b) Para cada $x \in K$ existe un abierto W_x de clausura compacta tal que $x \in W_x \subset \overline{W}_x \subset V$. Los abiertos W_x cubren a K. Tomamos un subcubrimiento finito y llamamos W a la unión de sus miembros. ■

Si X es un espacio localmente compacto y $f : X \longrightarrow \mathbb{R}$, llamaremos *soporte* de f a la clausura del conjunto de puntos donde f toma valores $\neq 0$. Llamaremos $C_c(X)$ al conjunto de las aplicaciones continuas $f : X \longrightarrow \mathbb{R}$ con soporte compacto. Es claro que se trata de un subespacio vectorial de $C(X)$.

Usaremos las notaciones $K \prec f$ y $f \prec V$ para indicar que $f : X \longrightarrow [0, 1]$, $f \in C_c(X)$, K es compacto, V es abierto, f toma el valor 1 en K y f toma el valor 0 en $X \setminus V$.

Teorema 7.26 (Lema de Urysohn) *Sea X un espacio localmente compacto, sea V un abierto y $K \subset V$ compacto. Entonces existe $f \in C_c(X)$ tal que $K \prec f \prec V$.*

DEMOSTRACIÓN: Por el teorema anterior existe un abierto W_0 de clausura compacta tal que $K \subset W_0 \subset \overline{W}_0 \subset V$. Sea $W_1 = X$. Aplicamos de nuevo el teorema a $\overline{W}_0 \subset V$, con lo que obtenemos un abierto $W_{1/2}$ de clausura compacta tal que
$$K \subset W_0 \subset \overline{W}_0 \subset W_{1/2} \subset \overline{W}_{1/2} \subset V \subset W_1.$$

Dos nuevas aplicaciones del mismo teorema nos dan
$$K \subset W_0 \subset \overline{W}_0 \subset W_{1/4} \subset \overline{W}_{1/4} \subset W_{1/2} \subset \overline{W}_{1/2} \subset W_{3/4} \subset \overline{W}_{3/4} \subset V \subset W_1.$$

Inductivamente vamos obteniendo una familia de abiertos $\{W_r\}_{r \in R}$, donde $R = \{k/2^i \mid i \in \mathbb{N}, 0 \leq k \leq 2^i\}$, de manera que si $r < r' < 1$ son puntos de R, entonces $K \subset \overline{W}_r \subset W_{r'} \subset V$.

Definimos $g : X \longrightarrow [0, 1]$ mediante $g(x) = \text{ínf}\{r \in R \mid x \in W_r\}$. Así $g[K] = \{0\}$ porque $K \subset W_0$ y $g[X \setminus V] = \{1\}$ porque $U_r \cap (X \setminus V) = \varnothing$ si $r < 1$. Basta probar que g es continua, pues entonces $f = 1 - g$ cumple lo pedido.

Sea $x \in X$ y $\epsilon > 0$. Si $g(x) \neq 0$ y $g(x) \neq 1$, entonces existen $r, r' \in R$ tales que $g(x) - \epsilon < r < g(x) < r' < g(x) + \epsilon$, luego $U = W_{r'} - \overline{W}_r$ es un entorno de x que cumple
$$g[U] \subset [r, r'] \subset \left]g(x) - \epsilon, g(x) + \epsilon\right[.$$

Si $g(x) = 0$ tomamos $0 = g(x) < r < g(x) + \epsilon$ y $U = W_r$ cumple lo mismo. Si $g(x) = 1$ tomamos $g(x) - \epsilon < r < g(x)$ y el U buscado es $X \setminus \overline{W}_r$. En cualquier caso obtenemos la continuidad de g en x. ■

Teorema 7.27 *Si X es un espacio localmente compacto, $V_1, \ldots, V_n$ son abiertos de X y $K \subset V_1 \cup \cdots \cup V_n$ es compacto, entonces existen funciones $h_i \prec V_i$ tales que $h_1(x) + \cdots + h_n(x) = 1$ para todo $x \in K$.*

Se dice que las funciones h_i forman una *partición de la unidad* subordinada a los abiertos dados.

Demostración: Dado $x \in K$ existe un i tal que $x \in V_i$. Existe un abierto W_x de clausura compacta tal que $x \in W_x \subset \overline{W}_x \subset V_i$. Los abiertos W_x cubren a K. Extraemos un subcubrimiento finito y llamamos H_i a la unión de todos los abiertos del subcubrimiento cuya clausura está en V_i. De este modo los H_i son abiertos de clausura compacta que cubren a K y $\overline{H}_i \subset V_i$. Por el teorema anterior existen funciones $\overline{H}_i \prec g_i \prec V_i$. Definimos

$$h_1 = g_1, \quad h_2 = (1 - g_1)g_2, \quad \ldots \quad h_n = (1 - g_1)(1 - g_2) \cdots (1 - g_{n-1})g_n.$$

Es claro que $h_i \prec V_i$ y una simple inducción prueba que

$$h_1 + \cdots + h_n = 1 - (1 - g_1) \cdots (1 - g_n).$$

Es claro entonces que la suma vale 1 sobre los puntos de K, pues una de las funciones g_i ha de tomar el valor 1. ■

Ya estamos en condiciones de demostrar el teorema de Riesz.

Teorema 7.28 (Teorema de representación de Riesz) *Sea X un espacio localmente compacto y sea $T : C_c(X) \longrightarrow \mathbb{R}$ una aplicación lineal tal que si $f \geq 0$ entonces $T(f) \geq 0$. Entonces existe una única medida de Borel casi regular μ en X tal que para toda función $f \in C_c(X)$ se cumple*

$$T(f) = \int_X f \, d\mu.$$

Demostración: Veamos primero la unicidad. Es claro que una medida casi regular está completamente determinada por los valores que toma sobre los conjuntos compactos, luego basta probar que si μ_1 y μ_2 representan a T en el sentido del teorema entonces $\mu_1(K) = \mu_2(K)$ para todo compacto K.

Por la regularidad existe un abierto V tal que $K \subset V$ y $\mu_2(V) < \mu_2(K) + \epsilon$. Por el lema de Urysohn existe una función $K \prec f \prec V$. Entonces

$$\begin{aligned} \mu_1(K) &= \int_X \chi_K \, d\mu_1 \leq \int_X f \, d\mu_1 = T(f) = \int_X f \, d\mu_2 \\ &\leq \int_X \chi_V \, d\mu_2 = \mu_2(V) < \mu_2(K) + \epsilon. \end{aligned}$$

Por consiguiente $\mu_1(K) \leq \mu_2(K)$ e igualmente se prueba la desigualdad contraria.

Para cada abierto V de X definimos $\mu(V) = \sup\{T(f) \mid f \prec V\}$. Es obvio que si $V_1 \subset V_2$ entonces $\mu(V_1) \leq \mu(V_2)$, luego si definimos

$$\mu(E) = \text{ínf}\{\mu(V) \mid E \subset V, \quad V \text{ abierto}\}, \quad \text{para todo } E \subset X,$$

es claro que la medida de un abierto es la misma en los dos sentidos en que la tenemos definida.

Aunque hemos definido la medida de cualquier conjunto, ésta sólo cumplirá las propiedades de las medidas al restringirla a una cierta σ-álgebra que contiene a la σ-álgebra de Borel. Concretamente, definimos $\mathcal{M}_F$ como la familia de subconjuntos E de X tales que $\mu(E) < +\infty$ y

$$\mu(E) = \sup\{\mu(K) \mid K \subset E, \quad K \text{ compacto}\}.$$

Definimos $\mathcal{M}$ como la familia de todos los $E \subset X$ tales que $E \cap K \in \mathcal{M}_F$ para todo compacto K. Probaremos que $\mathcal{M}$ es una σ-álgebra que contiene a la σ-álgebra de Borel y que la restricción de μ a $\mathcal{M}$ es una medida casi regular. Veremos también que $\mathcal{M}_F$ está formada por los conjuntos de $\mathcal{M}$ de medida finita. Dividimos la prueba en varios pasos.

1) *Si $\{E_i\}_{i=1}^{\infty}$ son subconjuntos de X, entonces*

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \mu(E_i).$$

Probamos primero que si V_1 y V_2 son abiertos $\mu(V_1 \cup V_2) \leq \mu(V_1) + \mu(V_2)$. Tomemos $g \prec V_1 \cup V_2$ arbitraria. Por el teorema 7.27 existen funciones h_1 y h_2 tales que $h_i \prec V_i$ y $h_1 + h_2$ vale 1 sobre los puntos del soporte de g. Por lo tanto $h_i g \prec V_i$, $g = h_1 g + h_2 g$.

$$T(g) = T(h_1 g) + T(h_2 g) \leq \mu(V_1) + \mu(V_2).$$

Como esto se cumple para toda $g \prec V_1 + V_2$, concluimos la desigualdad buscada.

Podemos suponer que $\mu(E_i) < +\infty$ para todo i, o la desigualdad que queremos probar se cumpliría trivialmente. Dado $\epsilon > 0$ la definición de μ implica que existen abiertos V_i que contienen a E_i de modo que $\mu(V_i) < \mu(E_i) + \epsilon/2^i$. Sea V la unión de todos los V_i y tomemos $f \prec V$. Como f tiene soporte compacto en realidad $f \prec V_1 \cup \cdots \cup V_n$ para algún n, luego

$$T(f) \leq \mu(V_1 \cup \cdots \cup V_n) \leq \mu(V_1) + \cdots + \mu(V_n) \leq \sum_{i=1}^{\infty} \mu(E_i) + \epsilon.$$

Como esto vale para toda $f \prec V$, resulta que

$$\mu(V) \leq \sum_{i=1}^{\infty} \mu(E_i) + \epsilon.$$

Ahora bien, la unión de los E_i está contenida en V, y la función μ es claramente monótona por su definición, luego

$$\mu\left(\bigcup_{i=1}^{\infty} E_i\right) \leq \sum_{i=1}^{\infty} \mu(E_i) + \epsilon.$$

Como esto vale para todo ϵ tenemos la desigualdad buscada.

2) *Si K es compacto, entonces $K \in \mathcal{M}_F$ y $\mu(K) = \operatorname{ínf}\{T(f) \mid K \prec f\}$. En particular los compactos tienen medida finita.*

   Si $K \prec f$ y $0 < \alpha < 1$, sea $V_\alpha = \{x \in X \mid f(x) > \alpha\}$. Entonces $K \subset V_\alpha$ y si $g \prec V_\alpha$ se cumple $\alpha g \leq f$. Por lo tanto

$$\mu(K) \leq \mu(V_\alpha) = \sup\{T(g) \mid g \prec V_\alpha\} \leq \alpha^{-1} T(f).$$

   Si hacemos que α tienda a 1 concluimos que $\mu(K) \leq T(f)$ y es obvio que K está en $\mathcal{M}_F$.

   Dado $\epsilon > 0$ existe un abierto V tal que $K \subset V$ y $\mu(V) < \mu(K) + \epsilon$. Existe una función $K \prec f \prec V$, luego

$$\mu(K) \leq T(f) \leq \mu(V) < \mu(K) + \epsilon,$$

   lo que prueba que $\mu(K) = \operatorname{ínf}\{T(f) \mid K \prec f\}$.

3) *$\mathcal{M}_F$ contiene a todos los abiertos de medida finita.*

   Sea V un abierto de medida finita y α un número real tal que $\alpha < \mu(V)$. Existe $f \prec V$ tal que $\alpha < T(f)$. Entonces $T(f) \leq \mu(K)$, pues en caso contrario existiría un abierto W tal que $K \subset W$ y $\mu(K) \leq \mu(W) < T(f)$, pero $f \prec W$, luego $T(f) \leq \mu(W)$, contradicción. Así hemos encontrado un compacto $K \subset V$ tal que $\alpha < \mu(K)$, lo que prueba que $V \in \mathcal{M}_F$.

4) *Si $\{E_i\}_{i=1}^{\infty}$ son elementos disjuntos de $\mathcal{M}_F$, y $E = \bigcup_{i=1}^{\infty} E_i$ entonces*

$$\mu(E) = \sum_{i=1}^{\infty} \mu(E_i).$$

   Si además $\mu(E) < +\infty$ entonces $E \in \mathcal{M}_F$.

   Veamos primero que si K_1 y K_2 son compactos disjuntos $\mu(K_1 \cup K_2) = \mu(K_1) + \mu(K_2)$. Dado $\epsilon > 0$ existe $K_1 \prec f \prec X \setminus K_2$. Por el paso 2) existe g tal que $K_1 \cup K_2 \prec g$ y $T(g) < \mu(K_1 \cup K_2) + \epsilon$. Claramente $K_1 \prec fg$ y $K_2 \prec (1-f)g$, luego

$$\mu(K_1) + \mu(K_2) \leq T(fg) + T(g - fg) = T(g) < \mu(K_1 \cup K_2) + \epsilon,$$

luego tenemos $\mu(K_1) + \mu(K_2) \leq \mu(K_1 \cup K_2)$ y el paso 1) nos da la otra desigualdad.

Pasando al caso general, de nuevo por 1) basta probar una desigualdad. Ésta es trivial si $\mu(E) = +\infty$, luego podemos suponer que E tiene medida finita. Fijado $\epsilon > 0$, puesto que $E_i \in \mathcal{M}_F$ existen compactos $H_i \subset E_i$ tales que $\mu(H_i) > \mu(E_i) - \epsilon/2^i$. Sea $K_n = H_1 \cup \cdots \cup H_n$. Entonces

$$\mu(E) \geq \mu(K_n) = \sum_{i=1}^{n} \mu(H_i) > \sum_{i=1}^{n} \mu(E_i) - \epsilon,$$

lo cual nos da claramente la desigualdad buscada. La desigualdad anterior muestra también que $\mu(K_n)$ tiende a $\mu(E)$ cuando n tiende a ∞ (una vez sabemos que la serie suma $\mu(E)$), lo que implica que $E \in \mathcal{M}_F$.

5) *Si $E \in \mathcal{M}_F$ y $\epsilon > 0$, existen un compacto K y un abierto V tales que $K \subset E \subset V$ y $\mu(V \setminus K) < \epsilon$.*

Por definición de $\mathcal{M}_F$ y de μ existen K y V tales que

$$\mu(V) - \frac{\epsilon}{2} < \mu(E) < \mu(K) + \frac{\epsilon}{2}.$$

Puesto que $V \setminus K$ es abierto, por 3) tenemos que $V \setminus K \in \mathcal{M}_F$, luego 4) implica que

$$\mu(K) + \mu(V \setminus K) = \mu(V) < \mu(K) + \epsilon.$$

6) *Si $A, B \in \mathcal{M}_F$ entonces $A \setminus B, A \cup B, A \cap B \in \mathcal{M}_F$.*

Aplicamos el paso anterior a los conjuntos A y B, lo que nos da conjuntos K_i y V_i, para $i = 1, 2$, de modo que $K_1 \subset A \subset V_1$, $K_2 \subset B \subset V_2$ y $\mu(V_i \setminus K_i) < \epsilon$. Entonces

$$A \setminus B \subset V_1 \setminus K_2 \subset (V_1 \setminus K_1) \cup (K_1 \setminus V_2) \cup (V_2 \setminus K_2),$$

luego el paso 1) implica $\mu(A \setminus B) \leq \epsilon + \mu(K_1 \setminus V_2) + \epsilon$ y $K_1 \setminus V_2$ es un subconjunto compacto de $A \setminus B$, luego esto prueba que $A \setminus B \in \mathcal{M}_F$.

Ahora, $A \cup B = (A \setminus B) \cup B$, luego el paso 4) implica que $A \cup B \in \mathcal{M}_F$ y como $A \cap B = A \setminus (A \setminus B)$, también $A \cap B \in \mathcal{M}_F$.

7) *$\mathcal{M}$ es una σ-álgebra que contiene a la σ-álgebra de Borel.*

Si $A \in \mathcal{M}$ y K es un compacto en X, entonces $(X \setminus A) \cap K = K \setminus (A \cap K)$, luego $(X \setminus A) \cap K$ es diferencia de dos elementos de $\mathcal{M}_F$, luego está en $\mathcal{M}_F$, luego $X \setminus A \in \mathcal{M}$.

Sea $A = \bigcup\limits_{i=1}^{\infty}$ una unión de elementos de $\mathcal{M}$. Si K es un compacto en X, tomamos $B_1 = A_1 \cap K$ y $B_n = (A_n \cap K) \setminus (B_1 \cup \cdots \cup B_{n-1}$, con lo que cada $B_n \in \mathcal{M}_F$ y son disjuntos dos a dos. Por 4) tenemos que $A \cap K$ (la unión de los B_n) está en $\mathcal{M}_F$, luego $A \in \mathcal{M}$. Esto prueba que $\mathcal{M}$ es una σ-álgebra.

Si C es un cerrado de X y K es compacto, entonces $C \cap K$ es compacto, luego está en $\mathcal{M}_F$, luego $C \in \mathcal{M}$. Por lo tanto $\mathcal{M}$ contiene a todos los cerrados, luego a todos los abiertos, luego a todos los conjuntos de Borel.

8) *$\mathcal{M}_F$ está formado por los conjuntos de $\mathcal{M}$ de medida finita.*

Si $E \in \mathcal{M}_F$, los pasos 2) y 6) implican que $E \cap K \in \mathcal{M}_F$, luego $E \in \mathcal{M}$. Recíprocamente, si $E \in \mathcal{M}$ tiene medida finita, dado $\epsilon > 0$ existe un abierto V que contiene a E y tiene medida finita. Por 3) y 5) existe un compacto $K \subset V$ con $\mu(V \setminus K) < \epsilon$. Como $E \cap K \in \mathcal{M}_F$, existe un compacto $H \subset E \cap K$ con $\mu(E \cap K) < \mu(H) + \epsilon$.

Puesto que $E \subset (E \cap K) \cup (V \setminus K)$, resulta

$$\mu(E) \leq \mu(E \cap K) + \mu(V \setminus K) < \mu(H) + 2\epsilon,$$

luego $E \in \mathcal{M}_F$.

Tras estas comprobaciones ya estamos en condiciones de probar el teorema. Consideremos la restricción de μ a la σ-álgebra de Borel. Los pasos 4) y 8) justifican que esta restricción es una medida. Hemos probado que μ es finita sobre los compactos, por definición se aproxima por abiertos y por 8) se aproxima por compactos en los conjuntos de medida finita. El argumento de 3) prueba de hecho que los abiertos de medida infinita contienen compactos de medida arbitrariamente grande, luego μ se aproxima por compactos en todos los abiertos. Así pues μ es casi regular.

Falta probar que μ representa a T. Dada $f \in C_c(X)$, basta probar que

$$T(f) \leq \int_X f \, d\mu,$$

pues aplicando esto mismo a $-f$ obtenemos la desigualdad opuesta. Sea K el soporte de f. Entonces $f[X] \subset f[K] \cup \{0\}$ es compacto, $f[X] \subset [a, b]$, para ciertos números reales a y b. Sea $\epsilon > 0$ y tomemos números

$$y_0 < a < y_1 < \cdots < y_n = b$$

tales que $y_{i+1} - y_i < \epsilon$.

Sea $E_i = \{x \in X \mid y_{i-1} < f(x) \leq y_i\} \cap K$. Como f es continua, f es medible respecto al álgebra de Borel, luego los conjuntos E_i son conjuntos de Borel disjuntos cuya unión es K. Existen abiertos V_i tales que $E_i \subset V_i$,

$$\mu(V_i) < \mu(E_i) + \frac{\epsilon}{n}$$

y $f(x) < y_i + \epsilon$ para todo $x \in V_i$. Por el teorema 7.27 existen funciones $h_i \prec V_i$ que suman 1 sobre K. Por lo tanto $f = h_1 f + \cdots h_n f$ y de 2) se sigue que

$$\mu(K) \leq T(h_1 + \cdots + h_n) = T(h_1) + \cdots + T(h_n).$$

Como $h_i f \leq (y_i+\epsilon)h_i$ e $y_i - \epsilon < f(x)$ en E_i, tenemos que

$$\begin{aligned}
T(f) &= \sum_{i=1}^n T(h_i f) \leq \sum_{i=1}^n (y_i+\epsilon)T(h_i) \\
&= \sum_{i=1}^n (|a|+y_i+\epsilon)T(h_i) - |a|\sum_{i=1}^n T(h_i) \\
&\leq \sum_{i=1}^n (|a|+y_i+\epsilon)(\mu(E_i)+\frac{\epsilon}{n}) - |a|\mu(K) \\
&= \sum_{i=1}^n (y_i-\epsilon)\mu(E_i) + 2\epsilon\mu(K) + \frac{\epsilon}{n}\sum_{i=1}^n(|a|+y_i+\epsilon) \\
&\leq \int_X f\,d\mu + \epsilon(2\mu(K)+|a|+b+\epsilon).
\end{aligned}$$

Como ϵ es arbitrario, tenemos la desigualdad buscada. ■

Recordemos que el teorema 7.5 implica que si el espacio X es σ-compacto entonces las medidas que proporciona el teorema de Riesz son regulares. Una aplicación interesante del teorema de Riesz nos da que en espacios razonables (como $\mathbb{R}^n$) toda medida razonable es regular:

Teorema 7.29 *Sea X un espacio localmente compacto en el que todo abierto sea σ-compacto. Si μ es una medida de Borel en X tal que todo compacto tiene medida finita entonces μ es regular.*

DEMOSTRACIÓN: Definamos el operador $T : C_c(X) \longrightarrow \mathbb{R}$ dado por $T(f) = \int_X f\,d\mu$. Notemos que si f tiene soporte K y M es una cota de f en K entonces

$$|T(f)| \leq \int_X |f|\,d\mu \leq M\mu(K) < +\infty,$$

luego T está bien definido y claramente cumple las hipótesis del teorema de Riesz. Por lo tanto existe una medida de Borel regular ν tal que para toda función $f \in C_c(X)$ se cumple

$$\int_X f\,d\nu = \int_X f\,d\mu.$$

Basta probar que $\mu = \nu$. Si V es un abierto, por hipótesis lo podemos expresar como unión de compactos $\{K_n\}_{n=1}^\infty$. Tomemos funciones $K_n \prec f_n \prec V$ y sea $g_n = \text{máx}\{f_1,\ldots,f_n\}$. Entonces $g_n \in C_c(X)$ y es claro que la sucesión $\{g_n\}_{n=1}^\infty$ es monótona creciente y converge puntualmente a χ_V. Por el teorema de la convergencia monótona resulta que

$$\nu(V) = \lim_n \int_X g_n\,d\nu = \lim_n \int_X g_n\,d\mu = \mu(V).$$

Así pues, μ y ν coinciden en los abiertos.

Dado un conjunto de Borel B, cortándolo con una familia creciente de compactos cuya unión sea X podemos expresarlo como unión creciente de conjuntos de Borel de medida finita para μ y ν, luego basta probar que ambas medidas coinciden sobre los conjuntos de Borel de medida finita. Por la regularidad de ν es fácil ver que existen un abierto V y un compacto K tales que $K \subset B \subset V$ y $\nu(V \setminus K) < \epsilon$, para un ϵ prefijado. Como $V \setminus K$ es abierto, esta última igualdad vale también para μ. Por consiguiente:

$$\begin{aligned} \mu(B) &\leq \mu(V) = \nu(V) \leq \nu(B) + \epsilon, \\ \nu(B) &\leq \nu(V) = \mu(V) \leq \nu(B) + \epsilon, \end{aligned}$$

luego $|\mu(B) - \nu(B)| \leq \epsilon$ para todo $\epsilon > 0$. ■

Terminamos la sección con un teorema importante sobre aproximación de funciones medibles por funciones continuas.

Teorema 7.30 (Teorema de Lusin) *Sea μ una medida de Borel regular en un espacio localmente compacto X y $f : X \longrightarrow \mathbb{R}$ una función medible tal que $f = 0$ salvo en un conjunto de medida finita. Dado $\epsilon > 0$ existe una función $g \in C_c(X)$ tal que $f = g$ salvo en un conjunto de medida menor que ϵ. Además podemos exigir que*

$$\sup_{x \in X} |g(x)| \leq \sup_{x \in X} |f(x)|.$$

DEMOSTRACIÓN: Sea $A = \{x \in X \mid f(x) \neq 0\}$. Supongamos primero que A es compacto y que $0 \leq f \leq 1$. Sea $\{s_n\}_{n=1}^{\infty}$ la sucesión monótona creciente de funciones simples construida en el teorema 7.10. Definimos $t_1 = s_1$ y $t_n = s_n - s_{n-1}$ para $n > 1$. Entonces $s_n(x) = k_n(f(x))/2^n$ y es fácil ver que $k_n(f(x)) = 2k_{n-1}(f(x))$ o bien $k_n(f(x)) = 2k_{n-1}(f(x)) + 1$, con lo que $2^n t_n$ toma sólo los valores 0 y 1. En otros términos $s_n = \chi_{T_n}$, para ciertos conjuntos medibles $T_n \subset A$. Además

$$f(x) = \sum_{n=1}^{\infty} t_n(x), \quad \text{para todo } x \in X.$$

Sea V un abierto de clausura compacta que contenga a A. Por regularidad existen compactos K_n y abiertos V_n de manera que $K_n \subset T_n \subset V_n \subset V$ y $\mu(V_n \setminus K_n) < 2^{-n}\epsilon$. Sea $K_n \prec h_n \prec V_n$. Definimos

$$g(x) = \sum_{n=1}^{\infty} 2^{-n} h_n(x).$$

Por el criterio de Weierstrass tenemos que g es continua. Además su soporte está contenido en la clausura de V, luego es compacto. Como $2^{-n}h_n(x) = t_n(x)$ excepto en $V_n \setminus K_n$, tenemos que $f = g$ excepto en $\bigcup_n (V_n \setminus K_n)$, que es un conjunto de medida menor que ϵ.

Del caso anterior se deduce el teorema para el caso en que A es compacto y f está acotada. Para probar el caso general tomamos $B_n = \{x \in X \mid |f(x)| > n\}$.

Estos conjuntos forman una familia decreciente con intersección vacía y todos tienen medida finita, luego $\mu(B_n)$ tiende a 0 con n. La función f coincide con la función acotada $h = (1 - \chi_{B_n})f$ salvo en B_n y, por otra parte, podemos tomar un compacto $K \subset A$ tal que $\mu(A \setminus K)$ sea arbitrariamente pequeño, luego basta aproximar $h\chi_K$ por una función continua, lo cual es posible por la parte ya probada.

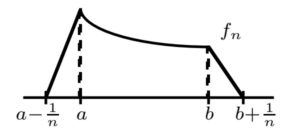
Por último, sea K el supremo de $|f|$. Si K es finito consideramos la función $h : \mathbb{R} \longrightarrow \mathbb{R}$ dada por

$$h(x) = \begin{cases} x & \text{si } |x| \leq K \\ \dfrac{Kx}{|x|} & \text{si } |x| > K \end{cases}$$

Claramente h es continua, $g_1 = g \circ h$ sigue cumpliendo el teorema y además su supremo no excede al de f. ■

7.5 La medida de Lebesgue

La medida de Lebesgue en $\mathbb{R}$ es la compleción de la medida que proporciona el teorema de Riesz cuando tomamos como aplicación lineal a la integral de Riemann. Así pues, la integral de Riemann y la integral de Lebesgue coinciden sobre las funciones continuas en $\mathbb{R}$ con soporte compacto. Es fácil ver que de hecho coinciden sobre toda función continua f sobre un intervalo $[a, b]$. Para ello extendemos f a una función f_n como indica la figura. Es claro que las funciones f_n son continuas en $\mathbb{R}$, tienen soporte compacto y convergen puntualmente a f. Además están dominadas por una función de la forma $c\chi_{[a-1,b+1]}$, para una constante c, luego por el teorema de la convergencia dominada tenemos que las integrales de las funciones f_n (Riemann y Lebesgue) tienden a la integral de lebesgue de f. Por otra parte, la integral de Riemann de f_n es la integral de Riemann de f más las integrales de Riemann en los intervalos $[a - 1/n, a]$ y $[b, b + 1/n]$, que están acotadas en módulo por una expresión de la forma K/n, luego tienden a 0. Así pues, las integrales de f_n convergen también a la integral de Riemann de f.



Puede probarse de hecho que todas las funciones integrables Riemann son integrables Lebesgue. Desde un punto de vista geométrico la integral de Lebesgue no aporta nada a la integral de Riemann, pues todas las funciones de interés son continuas o tienen un número finito de discontinuidades y siempre son integrables Riemann, pero desde un punto de vista técnico la integral de Lebesgue es mucho más potente. Los teoremas de convergencia y en general todos los resultados importantes sobre la integral de Lebesgue son falsos para la integral de Riemann.

Veamos ahora una construcción rápida de la integral de Riemann en $\mathbb{R}^n$ y de ella obtendremos la medida de Lebesgue en $\mathbb{R}^n$.

Definición 7.31 Una *celda* en $\mathbb{R}^n$ es un conjunto de la forma

$$W = \{x \in \mathbb{R}^n \mid \alpha_i < x_i < \beta_i,\ i = 1, \ldots, n\},$$

o cualquier conjunto obtenido reemplazando algunos de los signos $<$ por $\leq$ para algunos valores de i.

El *volumen* de una celda es

$$\operatorname{Vol}(W) = \prod_{i=1}^{n} (\beta_i - \alpha_i).$$

Si $a \in \mathbb{R}^n$ y $\delta > 0$ llamaremos *cubo* de vértice a y lado δ a la celda

$$C(a, \delta) = \{x \in \mathbb{R}^n \mid a_i \leq x_i < a_i + \delta\}.$$

Para cada natural $k > 0$ llamaremos P_k el conjunto de puntos de $\mathbb{R}^n$ cuyas coordenadas son de la forma $t2^{-k}$, para $t \in \mathbb{Z}$. Sea $\mathcal{C}_k$ el conjunto de todos los cubos de lado 2^{-k} con vértices en P_k. Sea $\mathcal{C}$ la unión de todos los conjuntos $\mathcal{C}_k$.

Teorema 7.32 *Se cumplen las propiedades siguientes:*

a) *$\mathbb{R}^n$ es la unión disjunta de los cubos de $\mathcal{C}_k$, para un k fijo.*

b) *Si $C \in \mathcal{C}_k$ y $C' \in \mathcal{C}_r$ con $r < k$ entonces $C \subset C'$ o $C \cap C' = \varnothing$.*

c) *Si $C \in \mathcal{C}_k$, entonces $\operatorname{Vol}(C) = 2^{kn}$ y si $r > k$ el conjunto P_r tiene exactamente $2^{(r-k)n}$ puntos en C.*

d) *Todo abierto no vacío de $\mathbb{R}^n$ es una unión numerable de cubos disjuntos de $\mathcal{C}$.*

Demostración: La única propiedad que no es obvia es la última. Si V es un abierto, es claro que cada $x \in V$ está en un cubo de $\mathcal{C}$ contenido en V. De entre todos los cubos de $\mathcal{C}$ contenidos en V tomamos todos los que están en $\mathcal{C}_1$, añadimos los de $\mathcal{C}_2$ que no están contenidos en ninguno de los de $\mathcal{C}_1$, añadimos los de $\mathcal{C}_3$ que no están contenidos en ninguno de los anteriores y así sucesivamente. El resultado es una familia numerable de cubos disjuntos de $\mathcal{C}$ cuya unión es V. ∎

Si $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ es una función con soporte compacto (no necesariamente continua) definimos

$$T_k(f) = 2^{-nk} \sum_{x \in P_k} f(x).$$

Claramente la suma tiene todos los sumandos nulos salvo un número finito de ellos. Si $f \in C_c(\mathbb{R}^n)$ tomamos una celda W que contenga a todos los cubos de $\mathcal{C}$ que cortan a su soporte. Como f es uniformemente continua (por 2.36), dado $\epsilon > 0$ existe un natural N tal que si $\|x-x'\|_\infty < 2^{-N}$ entonces $|f(x)-f(x')| < \epsilon$. Para cada $C \in \mathcal{C}_N$ sean m_C y M_C el ínfimo y el supremo de f en C (son todos nulos salvo un número finito de ellos). Sean g y h las funciones que en cada

cubo C toman respectivamente los valores m_C y M_C. Es claro entonces que $g \leq f \leq h$ y $h - g \leq \epsilon$. Si $k > N$, la propiedad c) del teorema anterior implica que

$$T_N(g) = T_k(g) \leq T_k(f) \leq T_k(h) = T_N(h),$$

luego los límites superior e inferior de la sucesión $\{T_k(f)\}_{k=1}^{\infty}$ difieren a lo sumo en $T_N(h-g) \leq \epsilon \mathrm{Vol}\,(W)$. Esto implica que existe

$$T(f) = \lim_k T_k(f).$$

La aplicación $T : C_c(\mathbb{R}^n) \longrightarrow \mathbb{R}$ así definida es obviamente lineal (se trata de la integral de Riemann) y también es claro que se encuentra en las hipótesis del teorema de Riesz.

Teorema 7.33 *Existe una única medida de Borel m en $\mathbb{R}^n$ tal que a cada celda le hace corresponder su volumen.*

DEMOSTRACIÓN: Tomamos como m la medida que proporciona el teorema de Riesz a partir del operador T que acabamos de construir. Consideremos una celda abierta W, es decir, un producto de intervalos abiertos. Sea E_k la unión de todos los cubos de $\mathcal{C}_k$ cuyas clausuras están contenidas en W. Éstos son un número finito y su unión E_k es una celda cuya clausura $\overline{E}$ está contenida en W. Tomemos $\overline{E}_k \prec f_k \prec W$ y sea $g_k = \max\{f_1, \ldots, f_k\}$. Es claro entonces que

$$\mathrm{Vol}\,(E_k) \leq T(f_k) \leq T(g_k) \leq \mathrm{Vol}\,(W).$$

También es fácil comprobar que cuando $k \to \infty$ el volumen de E_k tiende al de W, luego

$$\lim_n T(g_k) = \lim_n \int_X g_r\, dm = \mathrm{Vol}\,(W).$$

Por otro lado, $\{g_k\}_{k=1}^{\infty}$ es una sucesión creciente que converge puntualmente a χ_W, luego según el teorema de la convergencia monótona el límite anterior es también $m(W)$.

Una celda cerrada se expresa como intersección decreciente de celdas abiertas, luego las propiedades elementales de las medidas nos dan que m también asigna su volumen a cada celda cerrada. Puesto que m toma el mismo valor en una celda abierta que en su clausura, lo mismo vale para cualquier tipo de celda que esté comprendida entre ambas.

La unicidad es clara: dos medidas de Borel que asignen a cada celda su volumen toman valores finitos sobre los compactos (pues todo compacto está contenido en una celda), luego por el teorema 7.29 ambas son regulares. Puesto que todo abierto es unión numerable de cubos disjuntos (teorema 7.32), ambas medidas coinciden sobre los abiertos y por regularidad coinciden sobre cualquier conjunto de Borel. ■

Definición 7.34 La *medida de Lebesgue* en $\mathbb{R}^n$ es la compleción de la única medida de Borel que a cada celda le asigna su volumen. La representaremos por m.

La medida de Lebesgue se corresponde con el concepto geométrico de área y volumen en el caso de $\mathbb{R}^2$ y $\mathbb{R}^3$. En efecto, todas las figuras planas que aparecen en geometría (polígonos, elipses, etc.) tienen una frontera sin área, por lo que a efectos de calcular su área podemos considerar indistintamente su interior o su clausura. Si trabajamos con su interior, sabemos que puede expresarse como unión numerable de cubos disjuntos, por lo que el área debe ser la suma de las áreas de estos cubos, que es precisamente el valor de la medida de Lebesgue. Lo mismo vale para figuras tridimensionales. Es importante notar que todos los argumentos clásicos para el cálculo de áreas de figuras curvilíneas presuponen de un modo u otro este principio de que el área se puede calcular a partir de descomposiciones infinitas numerables.

Veamos ahora como se comporta la medida de Lebesgue frente a transformaciones afines.

Teorema 7.35 *Sea m la medida de Lebesgue en $\mathbb{R}^n$. Entonces:*

a) m es invariante por traslaciones, es decir, si $x \in \mathbb{R}^n$ y $A \subset \mathbb{R}^n$ es medible, entonces $x + A$ es medible y $m(x + A) = m(A)$.

b) Si $f : \mathbb{R}^n \longrightarrow \mathbb{R}^n$ es una aplicación lineal de determinante Δ y $A \subset \mathbb{R}^n$ es medible, entonces $f[A]$ es medible y $m\big(f[A]\big) = |\Delta| m(A)$.

Demostración: a) Es claro que el conjunto de los trasladados $x + B$ de los conjuntos de Borel forma una σ-álgebra que contiene a los abiertos, luego contiene a todos los conjuntos de Borel. Razonando igualmente con $-x$ concluimos que los trasladados de los conjuntos de Borel son exactamente los conjuntos de Borel. Si definimos $\mu(B) = m(x + B)$ tenemos claramente una medida de Borel y es fácil ver que a cada celda le asigna su volumen. Por consiguiente se trata de la medida de Lebesgue, es decir, tenemos que $m(x + B) = m(B)$ para todo conjunto de Borel B. Teniendo en cuenta la definición de la compleción es fácil extender este hecho a todo conjunto medible.

b) Supongamos primero que f es biyectiva. Puesto que también es continua, el mismo razonamiento que en el caso de las traslaciones implica que cuando B recorre los conjuntos de Borel de $\mathbb{R}^n$, lo mismo sucede con $f[B]$, así como que $\mu(B) = m\big(f[B]\big)$ define una medida de Borel regular. La linealidad de f y el apartado anterior implican que μ es invariante por traslaciones. Sea C un cubo cualquiera de lado 1 y sea $c = \mu(C)$. Si C' es un cubo de lado 2^{-k}, entonces C se expresa como unión disjunta de 2^{nk} trasladados de C', todos con la misma medida $\mu(C')$, luego $\mu(C') = c2^{-nk} = c\, m(C')$. Como todo abierto V es unión numerable de cubos disjuntos de lado 2^{-k}, concluimos que $\mu(V) = c\, m(V)$, y por regularidad lo mismo vale para todo conjunto de Borel B, es decir,

$$m\big(f[B]\big) = c\, m(B).$$

A partir de la definición de compleción se concluye que esta relación es cierta para todo conjunto medible Lebesgue.

Si f no es biyectiva entonces su imagen es un subespacio propio de $\mathbb{R}^n$. Basta probar que los subespacios propios de $\mathbb{R}^n$ tienen medida nula, con lo que

la relación anterior será cierta con $c = 0$. A través de un automorfismo, todo subespacio propio se transforma en un subespacio de $V = \{x \in \mathbb{R}^n \mid x_1 = 0\}$. Por la parte ya probada basta ver que V es nulo. Ahora bien, V es unión numerable de cubos degenerados de la forma $\{0\} \times [-k, k] \times \cdots \times [-k, k]$, luego basta ver que estos cubos son nulos, pero cada uno de ellos es intersección numerable de cubos $[-1/r, 1/r] \times [-k, k] \times \cdots \times [-k, k]$, cuya medida tiende a 0.

Falta probar que la constante c es exactamente $|\Delta|$. Si f no es biyectiva es evidente, pues hemos visto que $c = 0$. En caso contrario sabemos que c es la medida de la imagen por f de un cubo cualquiera de lado 1, por ejemplo $C = [0, 1[^n$.

Es conocido que todo automorfismo de $\mathbb{R}^n$ se descompone en producto de automorfismos que sobre la base canónica $\{e_1, \ldots, e_n\}$ actúan de una de las tres formas siguientes:

1) $\big(f(e_1), \ldots, f(e_n)\big)$ es una permutación de $(e_1, \ldots, e_n)$,

2) $f(e_1) = \alpha e_1$, $f(e_i) = e_i$, para $i = 2, \ldots, n$,

3) $f(e_1) = e_1 + e_2$, $f(e_i) = e_i$, para $i = 2, \ldots, n$.

Teniendo en cuenta además que el determinante de la composición de dos automorfismos es el producto de sus determinantes, es claro que basta probar el teorema cuando f es de uno de estos tipos.

Si f es del primer tipo tenemos $\Delta = \pm 1$ y $f[C] = C$, luego $c = 1$ y se cumple el teorema. Si f es del segundo tipo entonces $\Delta = \alpha$ y $f[C] = I \times [0, 1[^{n-1}$, donde $I = [0, \alpha[$ o $I =]\alpha, 0]$, según el signo de α, luego $c = |\alpha|$ como queríamos probar. En el último caso $\Delta = 1$ y $f[C]$ es el conjunto de los $x \in \mathbb{R}^n$ tales que

$$x_1 \leq x_2 < x_1 + 1, \quad 0 \leq x_i < 1 \quad \text{si } i \neq 2.$$

Sea S_1 el conjunto de $x \in f[C]$ con $x_2 < 1$ y $S_2 = f[C] \setminus S_1$. Entonces $C = S_1 \cup (S_2 - e_2)$, y la unión es disjunta. Por lo tanto

$$c = m(S_1 \cup S_2) = m(S_1) + m(S_2) = m(S_1) + m(S_2 - e_2) = m(C) = 1,$$

luego el teorema está probado. ■

Con esto hemos obtenido una interpretación del módulo del determinante de una aplicación lineal f. Por ejemplo, si $|\Delta| = 3$ esto significa que f triplica el volumen de los conjuntos. Recordemos que el signo del determinante ya lo interpretamos en el capítulo II.

De momento no estamos en condiciones de calcular integrales de funciones de varias variables. Los resultados que nos permitirán trabajar cómodamente con tales integrales los veremos en los próximos capítulos, especialmente en el siguiente. De momento vamos a ver un ejemplo interesante de integral en $\mathbb{R}$:

Definición 7.36 La *función factorial*[1] es la función $\Pi : \,]-1,+\infty[\, \longrightarrow \mathbb{R}$ dada por

$$\Pi(x) = \int_0^{+\infty} t^x e^{-t}\, dt.$$

Recordemos que $t^x = e^{x \log t}$. Vamos a probar que el integrando es realmente una función integrable en $]0,+\infty[$ para todo $x > -1$. Probamos por separado que es integrable en $]0,1]$ y en $[1,+\infty[$.

Si $x \geq 0$ entonces $t^x e^{-t}$ es una función continua en $[0,1]$, luego es integrable. Si $-1 < x < 0$ entonces (si $0 \leq t \leq 1$) se cumple $0 \leq t^x e^{-t} \leq t^x$, y la función t^x es integrable: su integral es

$$\int_0^1 t^x\, dt = \lim_n \int_{1/n}^1 t^x\, dt = \lim_n \left[\frac{t^{x+1}}{x+1}\right]_{1/n}^1 = \left[\frac{t^{x+1}}{x+1}\right]_0^1 = \frac{1}{x+1}.$$

La primera igualdad se sigue del teorema de la convergencia monótona. Siempre que tengamos una integral de una función no negativa definida sobre un intervalo y de modo que restringida a intervalos menores sea continua y acotada podemos aplicar esta técnica (aplicar la regla de Barrow en intervalos menores y tomar el límite). En situaciones similares pasaremos directamente del primer término al tercero sobrentendiendo el límite.

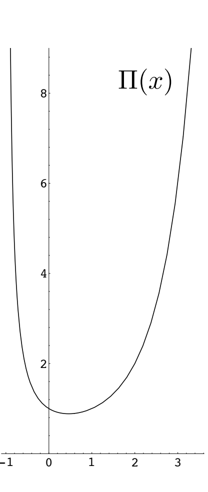
Consideremos ahora el intervalo $[1,+\infty[$. Observemos que

$$\lim_{t\to+\infty} \frac{t^x e^{-t}}{1/t^2} = \lim_{t\to+\infty} t^{x+2} e^{-t} = 0,$$

(acotamos $x+2$ por un número natural n y aplicamos n veces la regla de l'Hôpital.)

Esto implica que existe un $M > 0$ tal que si $t \geq M$ entonces $t^x e^{-t} \leq 1/t^2$. La función $t^x e^{-t}$ es continua en el intervalo $[1,M]$, luego es integrable, luego basta probar que también lo es en $[M,+\infty[$ y a su vez basta que lo sea $1/t^2$, pero

$$\int_M^{+\infty} \frac{1}{t^2}\, dt = \left[-\frac{1}{t}\right]_M^{+\infty} = \frac{1}{M}.$$



Con esto tenemos probada la existencia de la función Π. La figura muestra su gráfica. ■

El teorema siguiente recoge las propiedades más importantes de la función factorial, entre ellas la que le da nombre:

[1]La función factorial fue descubierta y estudiada por Euler, aunque fue Gauss quien la expresó en la forma que aquí usamos como definición. Legendre introdujo el cambio de variable $\Gamma(x) = \Pi(x-1)$, que inexplicablemente ha prevalecido sobre la notación de Gauss y actualmente la función es más conocida como "función Gamma". Nosotros respetamos la notación de Gauss pues, como veremos, resulta mucho más natural.

Teorema 7.37 *La función factorial*

$$\Pi(x) = \int_0^{+\infty} t^x e^{-t}\, dt.$$

es continua en $]-1, +\infty[$ *y cumple la ecuación funcional*

$$\Pi(x+1) = (x+1)\Pi(x).$$

Además $\Pi(0) = 1$*, de donde*[2] $\Pi(n) = n!$ *para todo número natural* n*.*

Demostración: Para probar que Π es continua basta probar que lo es en un intervalo $I =]-1+\epsilon, M[$. En este intervalo el integrando de Π está mayorado por la función $g(t) = t^{-1+\epsilon}e^{-t} + t^M e^{-t}$, que es integrable en $]0, +\infty[$, pues su integral es $\Pi(-1+\epsilon) + \Pi(M)$. Si llamamos

$$\Pi_n(x) = \int_{-1+1/n}^{n} t^x e^{-x}\, dt,$$

el teorema 7.23 garantiza que Π_n es continua en I y basta probar que Π_n converge uniformemente a Π en I. Ahora bien, si $x \in I$ tenemos que

$$|\Pi(x) - \Pi_n(x)| \leq \int_0^{1/n} g(t)\, dt + \int_n^{+\infty} g(t)\, dt,$$

y es claro que el segundo miembro tiende a 0 con n.

La ecuación funcional se obtiene integrando por partes. En efecto:

$$\Pi(x+1) = \int_0^{+\infty} t^{x+1}e^{-t}\, dt = \big[-t^{x+1}e^{-t}\big]_0^{+\infty} + (x+1)\int_0^{+\infty} t^x e^{-t}\, dt = (x+1)\Pi(x).$$

Claramente $\Pi(0) = \int_0^{+\infty} e^{-t}\, dt = [-e^{-t}]_0^{+\infty} = 1 = 0!$, luego por inducción tenemos la igualdad $\Pi(n) = n!$ ■

Puede probarse que Π es de clase C^∞ en su dominio. La función factorial tiene numerosas aplicaciones en análisis real y complejo, desde la estadística hasta la teoría de números. Nosotros no iremos más allá en su estudio, aunque más adelante veremos algún ejemplo adicional en torno a ella.

[2] Con la notación de Legendre las propiedades de la función factorial quedan distorsionadas. Por ejemplo, la función gamma cumple $\Gamma(x+1) = x\Gamma(x)$, $\Gamma(n) = (n-1)!$

Capítulo VIII

Teoría de la medida II

En este capítulo profundizaremos en la teoría de la medida hasta obtener los principales resultados necesarios para trabajar con integrales de funciones de varias variables. Los resultados más importantes serán el teorema de Fubini, que reduce el cálculo de la integral de una función de n variables al de n integrales sucesivas de funciones de una variable, y el teorema de cambio de variable, que generaliza al que ya conocemos para funciones de una variable (integración por sustitución).

8.1 Producto de medidas

Aquí vamos a definir un producto de medidas, de modo que, por ejemplo, la medida de Lebesgue en $\mathbb{R}^n$ será el producto de la medida de Lebesgue en $\mathbb{R}$ consigo misma n veces. Después probaremos el teorema de Fubini, que reduce el cálculo de una integral respecto a la medida producto al cálculo de integrales respecto a los factores.

Definición 8.1 Sean X e Y dos conjuntos y $\mathcal{A}$, $\mathcal{B}$ dos σ-álgebras de subconjuntos de X e Y respectivamente (a cuyos elementos llamaremos conjuntos medibles). Un *rectángulo medible* en $X \times Y$ es un conjunto de la forma $A \times B$, donde $A \in \mathcal{A}$ y $B \in \mathcal{B}$. Llamaremos *figuras elementales* a las uniones disjuntas de rectángulos medibles. Llamaremos $\mathcal{A} \times \mathcal{B}$ a la σ-álgebra generada por los rectángulos medibles. Cuando hablemos de conjuntos medibles en $X \times Y$ entenderemos que nos referimos a los de $\mathcal{A} \times \mathcal{B}$.

Si $E \subset X \times Y$, $x \in X$, $y \in Y$, definimos las *secciones* de E determinadas por x e y como

$$E_x = \{y \in Y \mid (x, y) \in E\}, \quad E^y = \{x \in X \mid (x, y) \in E\}.$$

Teorema 8.2 *En las condiciones anteriores, si E es medible en $X \times Y$, entonces E_x y E^y son medibles en Y y en X respectivamente.*

Demostración: Sea $\mathcal{C}$ el conjunto de todos los $E \in \mathcal{A}\times\mathcal{B}$ tales que $E_x \in \mathcal{B}$ para todo $x \in X$. Si $E = A \times B$ entonces $E_x = B$ para todo $x \in A$, y $E_x = \varnothing$ si $x \notin A$ luego todos los rectángulos medibles están en $\mathcal{C}$. Ahora vemos que $\mathcal{C}$ es una σ-álgebra, de donde se sigue que $\mathcal{A} \times \mathcal{B} \subset \mathcal{C}$, lo que prueba el teorema (para E_x, el caso E^y es análogo).

a) Obviamente $X \times Y \in \mathcal{C}$.

b) Si $E \in \mathcal{C}$, entonces $(X \times Y \setminus E)_x = Y \setminus E_x \in \mathcal{B}$, luego $X \times Y \setminus E \in \mathcal{C}$.

c) Si $\{E_i\}_{i=1}^{\infty} \subset \mathcal{C}$ y $E = \bigcup\limits_{i=1}^{\infty} E_i$, entonces $E_x = \bigcup\limits_{i=1}^{\infty} E_{ix} \in \mathcal{B}$. Por lo tanto $E \in \mathcal{C}$.

■

Vamos a dar una caracterización de $\mathcal{A} \times \mathcal{B}$ que nos será útil después. Para ello definimos una *clase monótona* $\mathcal{M}$ en un conjunto X como una colección de subconjuntos de X tal que si $\{A_i\}_{i=1}^{\infty}$ es una familia creciente de conjuntos de $\mathcal{M}$ (es decir, $A_i \subset A_{i+1}$) entonces $\bigcup\limits_{i=1}^{\infty} A_i \in \mathcal{M}$ y si la familia es decreciente $(A_{i+1} \subset A_i)$ entonces $\bigcap\limits_{i=1}^{\infty} A_i \in \mathcal{M}$.

Es claro que la intersección de clases monótonas es de nuevo una clase monótona, por lo que podemos hablar de la clase monótona generada por un conjunto, es decir, la menor clase monótona que lo contiene.

Teorema 8.3 *En las condiciones anteriores, $\mathcal{A} \times \mathcal{B}$ es la clase monótona generada por las figuras elementales.*

Demostración: Sea $\mathcal{M}$ la clase monótona generada por las figuras elementales. Claramente $\mathcal{A} \times \mathcal{B}$ es una clase monótona que contiene a las figuras elementales, luego $\mathcal{M} \subset \mathcal{A} \times \mathcal{B}$. Las igualdades

$$\begin{aligned}(A_1 \times B_1) \cap (A_2 \times B_2) &= (A_1 \cap A_2) \times (B_1 \times B_2)\\ (A_1 \times B_1) \setminus (A_2 \times B_2) &= \big((A_1 \setminus A_2) \times B_1\big) \cup \big((A_1 \cap A_2) \times (B_1 \setminus B_2)\big)\end{aligned}$$

muestran que la intersección de dos rectángulos medibles es un rectángulo medible y que su diferencia es la unión de dos rectángulos medibles disjuntos, luego una figura elemental. De aquí se sigue claramente que la intersección y la diferencia de figuras elementales es una figura elemental. Lo mismo vale para la unión, pues $P \cup Q = (P \setminus Q) \cup Q$, y la unión es disjunta.

Para cada conjunto $P \subset X \times Y$ definimos $\mathcal{M}_P$ como el conjunto de todos los $Q \subset X \times Y$ tales que $P \setminus Q \in \mathcal{M}$, $Q \setminus P \in \mathcal{M}$ y $P \cup Q \in \mathcal{M}$. Las propiedades siguientes son obvias:

1) $Q \in \mathcal{M}_P$ si y sólo si $P \in \mathcal{M}_Q$.

2) $\mathcal{M}_P$ es una clase monótona.

Si P es una figura elemental, hemos probado que toda figura elemental está contenida en $\mathcal{M}_P$, de donde 2) implica que $\mathcal{M} \subset \mathcal{M}_P$. Fijemos ahora $Q \in \mathcal{M}$. Si P es una figura elemental, por 1) tenemos que $P \in \mathcal{M}_Q$, luego por 2) resulta que $\mathcal{M} \subset \mathcal{M}_Q$.

Con esto hemos probado que la diferencia y la unión de dos elementos de $\mathcal{M}$ está en $\mathcal{M}$. Al añadir esto a la monotonía concluimos que $\mathcal{M}$ es una σ-álgebra. En efecto, ciertamente $X \times Y$ está en $\mathcal{M}$, luego el complemento de un elemento de $\mathcal{M}$ está en $\mathcal{M}$. Si $\{A_i\}_{i=1}^\infty$ es una familia de elementos de $\mathcal{M}$, entonces las uniones $B_i = A_1 \cup \cdots \cup A_i$ están en $\mathcal{M}$, pero la unión de los A_i es la misma que la de los B_i, que está en $\mathcal{M}$ por monotonía.

Puesto que las figuras elementales están en $\mathcal{M}$, concluimos que $\mathcal{A} \times \mathcal{B} = \mathcal{M}$. ■

Veamos ahora la relación entre la medibilidad de funciones en un producto y en los factores. Siempre en las mismas condiciones, si $f : X \times Y \longrightarrow Z$, $x \in X$, $y \in Y$, definimos $f_x : Y \longrightarrow Z$ y $f^y : X \longrightarrow Z$ como las aplicaciones dadas por $f_x(y) = f(x, y)$, $f^y(x) = f(x, y)$.

Teorema 8.4 *Si $f : X \times Y \longrightarrow Z$ es una función medible, entonces f_x es medible para todo $x \in X$ y f^y es medible para todo $y \in Y$.*

Demostración: Si V es un abierto en Z, claramente $f_x^{-1}[V] = f^{-1}[V]_x$, luego es medible. Por lo tanto f_x es medible. Igualmente se razona con f^y. ■

Con esto estamos casi a punto de definir el producto de medidas. La definición se apoyará en el teorema siguiente.

Teorema 8.5 *Sean X e Y espacios medida con medidas σ-finitas μ y ν. Sea E un subconjunto medible de $X \times Y$. Entonces las aplicaciones $\nu(E_x)$ y $\mu(E^y)$ son funciones medibles de x e y respectivamente. Además*

$$\int_X \nu(E_x)\, d\mu = \int_Y \mu(E^y)\, d\nu.$$

Demostración: Notar que por el teorema 8.2 los conjuntos E_x, E^y son medibles, luego tiene sentido considerar $\nu(E_x)$ y $\mu(E^y)$.

Llamemos $\mathcal{C}$ a la familia de todos los subconjuntos medibles de $X \times Y$ para los que se cumple el teorema. Vamos a probar que $\mathcal{C}$ tiene las propiedades siguientes:

a) $\mathcal{C}$ contiene a los rectángulos medibles.

b) Si $\{Q_n\}_{n=1}^\infty \subset \mathcal{C}$ es creciente entonces $Q = \bigcup_{n=1}^\infty Q_n \in \mathcal{C}$.

c) Si $\{Q_n\}_{n=1}^\infty \subset \mathcal{C}$ son disjuntos dos a dos entonces $Q = \bigcup_{n=1}^\infty Q_n \in \mathcal{C}$.

d) Si $\{Q_n\}_{n=1}^\infty \subset \mathcal{C}$ es decreciente y $Q_1 \subset U \times V$, con $\mu(U), \nu(V) < +\infty$, entonces $Q = \bigcap_{n=1}^\infty Q_n \in \mathcal{C}$.

En efecto, si $U \times V$ es un rectángulo medible, entonces

$$(U \times V)_x = \begin{cases} V & \text{si } x \in U \\ \varnothing & \text{si } x \notin U. \end{cases}$$

Por lo tanto $\nu\big((U \times V)_x\big) = \nu(V)\chi_U$, que es una función medible. Igualmente $\mu\big((U \times V)_y\big) = \mu(U)\chi_V$. Las integrales valen ambas $\mu(U)\nu(V)$, luego $U \times V$ está en $\mathcal{C}$. Esto prueba a).

Para demostrar b) observamos que $Q_x = \bigcup\limits_{n=1}^{\infty} (Q_n)_x$, y la sucesión es creciente, por lo que $\nu(Q_x) = \lim\limits_n \nu\big((Q_n)_x\big)$. Como los conjuntos Q_n están en $\mathcal{C}$, las funciones $\nu\big((Q_n)_x\big)$ son medibles, luego su límite puntual $\nu(Q_x)$ también lo es. Igualmente ocurre con $\mu(Q_y)$. El teorema de la convergencia monótona da la igualdad de las integrales, luego $Q \in \mathcal{C}$.

La prueba de c) es similar, usando ahora que $\nu(Q_x)$ es la suma de las funciones $\nu\big((Q_n)_x\big)$ en lugar del límite.

En el caso d) tenemos también $\nu(Q_x) = \lim\limits_n \nu\big((Q_n)_x\big)$, pero ahora la sucesión no es monótona creciente. La única diferencia es que en lugar del teorema de la convergencia monótona usamos el teorema de la convergencia dominada. La hipótesis $\nu(V) < +\infty$ garantiza que las funciones $\nu\big((Q_n)_x\big)$ están dominadas por la función integrable χ_V.

Estamos suponiendo que las medidas en X y en Y son σ-finitas, lo cual significa que podemos expresar $X = \bigcup\limits_{n=1}^{\infty} X_n$ e $Y = \bigcup\limits_{n=1}^{\infty} Y_n$ para ciertos conjuntos medibles de medida finita que además podemos suponer disjuntos dos a dos.

Sea ahora E un conjunto medible en $X \times Y$. Definamos $E_{mn} = E \cap (X_m \times Y_n)$ y sea $\mathcal{M}$ la familia de todos los conjuntos E tales que los E_{mn} así definidos están en $\mathcal{C}$. las propiedades b) y d) muestran que $\mathcal{M}$ es una clase monótona, mientras que a) y c) muestran que contiene a las figuras elementales. El teorema 8.3 implica ahora que $\mathcal{M}$ contiene a todos los conjuntos medibles de $X \times Y$.

Así pues, para todo conjunto medible E, los conjuntos E_{mn} están en $\mathcal{C}$, pero claramente E es unión disjunta de los E_{mn},luego por c) tenemos $E \in \mathcal{C}$, es decir, todo conjunto medible E cumple el teorema. ■

Definición 8.6 Sean X e Y espacios con medidas σ-finitas μ y ν. Definimos la *medida producto* $\mu \times \nu$ como la dada por

$$(\mu \times \nu)(Q) = \int_X \nu(Q_x)\, d\mu(x) = \int_Y \mu(Q^y)\, d\nu(y).$$

Con el teorema 7.16 se prueba fácilmente que $\mu \times \nu$ es realmente una medida en la σ-álgebra producto. Además es claro que sobre los rectángulos medibles tenemos $(\mu \times \nu)(A \times B) = \mu(A)\nu(B)$ (con el convenio $0 \cdot \infty = 0$).

Conviene dar una caracterización de la medida producto que no dependa del teorema anterior:

Teorema 8.7 *Dados dos espacios X e Y con medidas σ-finitas μ y ν, la medida producto es la única que cumple que $(\mu \times \nu)(A \times B) = \mu(A)\nu(B)$ para todo rectángulo medible $A \times B$.*

Demostración: Supongamos que dos medidas λ_1 y λ_2 se comportan sobre los rectángulos medibles como la medida producto. Descompongamos $X = \bigcup_{n=1}^{\infty} X_n$ e $Y = \bigcup_{n=1}^{\infty} Y_n$, para ciertos conjuntos medibles de medida finita disjuntos dos a dos. Sea $\mathcal{M}$ la familia de los conjuntos medibles E de $X \times Y$ tales que $\lambda_1\big(E \cap (X_m \times Y_n)\big) = \lambda_2\big(E \cap (X_m \times Y_n)\big)$ para todo m, n. Es claro que $\mathcal{M}$ es una clase monótona que contiene a las figuras elementales, luego por el teorema 8.3 tenemos que $\mathcal{M}$ contiene a todos los conjuntos medibles. De aquí se sigue que las dos medidas coinciden sobre cualquier conjunto medible. ■

Veamos ahora que toda esta teoría es aplicable a la medida de Lebesgue en $\mathbb{R}^n$. Primero probemos un hecho general:

Teorema 8.8 *Si X e Y son dos espacios topológicos con bases numerables, entonces el producto de las σ-álgebras de Borel es la σ-álgebra de Borel del producto. En particular el producto de medidas de Borel es una medida de Borel.*

Demostración: Si U y V son conjuntos de Borel en X e Y respectivamente, entonces $U \times Y$ es un conjunto de Borel en el producto, pues es la antiimagen de U por la proyección en X, que es continua, luego medible. Igualmente $X \times V$ es un conjunto de Borel, y también lo es $U \times V$ por ser la intersección de ambos. De aquí se sigue que todas las figuras elementales son conjuntos de Borel, luego también lo son todos los conjuntos medibles en $X \times Y$. Recíprocamente, los productos de abiertos básicos $U \times V$ forman una base numerable de $X \times Y$, luego todo abierto de $X \times Y$ es unión numerable de estos abiertos básicos, luego todo abierto de $X \times Y$ es medible, luego todo conjunto de Borel es medible. ■

Teorema 8.9 *La medida de Lebesgue en $\mathbb{R}^{m+n}$ (restringida a los conjuntos de Borel) es el producto de la medida de Lebesgue en $\mathbb{R}^m$ por la medida de Lebesgue en $\mathbb{R}^n$ (restringidas ambas a los conjuntos de Borel).*

En efecto, es claro que las celdas son rectángulos medibles, y la medida producto coincide sobre ellas con la medida de Lebesgue, luego es la medida de Lebesgue. ■

Ahora probamos el teorema principal de esta sección:

Teorema 8.10 (Teorema de Fubini) *Sean X e Y dos espacios medida con medidas σ-finitas μ y ν y sea $f : X \times Y \longrightarrow [-\infty, +\infty]$ una función medible.*

a) Si $f \geq 0$, entonces las funciones

$$\int_Y f_x \, d\nu \quad \textit{(como función de } x\textit{)} \quad e \quad \int_X f_y \, d\mu \quad \textit{(como función de } y\textit{)}$$

son medibles, y se cumple

$$\int_{X\times Y} f\, d(\mu\times\nu) = \int_X \left(\int_Y f_x\, d\nu\right) d\mu = \int_Y \left(\int_X f_y\, d\mu\right) d\nu.$$

b) Si $\int_X \left(\int_Y |f|_x\, d\nu\right) d\mu < +\infty$, *entonces* $f \in L^1(\mu\times\nu)$.

c) Si $f \in L^1(\mu\times\nu)$, *entonces* $f_x \in L^1(\nu)$ *p.c.t.* y, $f_y \in L^1(\mu)$ *p.c.t.* x, *y las funciones definidas en a) p.c.t.p. están en* $L^1(\mu)$ *y* $L^1(\nu)$ *respectivamente y sus integrales coinciden (según se afirma en a).*

Demostración: a) Por el teorema 8.4, las funciones f_x y f_y son medibles, luego tienen sentido sus integrales. Si $f = \chi_Q$, para un cierto conjunto medible $Q \subset X\times Y$, entonces a) se reduce al teorema 8.5 y a la definición de la medida producto. De aquí se sigue que las igualdades de a) son válidas cuando f es una función simple. En general existe una sucesión creciente de funciones simples $\{s_n\}_{n=1}^\infty$ que converge puntualmente a f. Entonces es claro que $\{(s_n)_x\}_{n=1}^\infty$ converge puntualmente a f_x y, por el teorema de la convergencia monótona concluimos que

$$\lim_n \int_Y (s_n)_x\, d\nu = \int_Y f_x\, d\nu.$$

Como las funciones simples cumplen a) tenemos que las funciones $\int_Y (s_n)_x\, d\nu$ son medibles y

$$\int_{X\times Y} s_n\, d(\mu\times\nu) = \int_X \left(\int_Y (s_n)_x\, d\nu\right) d\mu.$$

Consecuentemente el límite $\int_Y f_x\, d\nu$ es medible y aplicando el teorema de la convergencia monótona a los dos miembros de la igualdad anterior queda

$$\int_{X\times Y} f\, d(\mu\times\nu) = \int_X \lim_n \left(\int_Y (s_n)_x\, d\nu\right) d\mu = \int_X \left(\int_Y f_x\, d\nu\right) d\mu.$$

La otra igualdad se prueba análogamente.

Las hipótesis de b) implican por a) que $|f|$ es integrable, luego f también lo es.

Para probar c) descompongamos $f = f^+ - f^-$. Tenemos que f^+ y f^- son integrables. Por a) la integrabilidad de f^+ significa que

$$\int_{X\times Y} f^+ d(\mu\times\nu) = \int_X \left(\int_Y f_x^+\, d\nu\right) d\mu < +\infty,$$

luego $\int_Y f_x^+\, d\nu$ ha de ser finita salvo a lo sumo en un conjunto nulo, y lo mismo el válido para $\int_Y f_x^-\, d\nu$. Salvo para los puntos x en la unión de los dos conjuntos nulos, tenemos que la integral

$$\int_Y f_x\, d\nu = \int_Y f_x^+\, d\nu - \int_Y f_x^-\, d\nu$$

está definida y es finita, es decir, que la función $\int_Y f_x \, d\nu$ es integrable. Además

$$\int_{X\times Y} f \, d(\mu\times\nu) = \int_{X\times Y} f^+ \, d(\mu\times\nu) - \int_{X\times Y} f^- \, d(\mu\times\nu)$$

$$= \int_X \left(\int_Y f_x^+ \, d\nu\right) d\mu - \int_X \left(\int_Y f_x^- \, d\nu\right) d\mu = \int_X \left(\int_Y f_x \, d\nu\right) d\mu.$$

La otra parte de c) es análoga. ■

En el caso de la medida de Lebesgue en $\mathbb{R}^n$ sustituiremos dm por $dx_1 \cdots dx_n$. En estos términos el teorema de Fubini (por ejemplo para dos variables) se expresa como sigue:

$$\int_{\mathbb{R}^2} f(x,y) \, dxdy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(x,y) \, dx\right) dy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} f(x,y) \, dy\right) dx.$$

Si trabajamos con funciones continuas sobre un compacto no hemos de preocuparnos de la integrabilidad.

Ejemplo Vamos a calcular el área de la elipse E de semiejes a, y b, formada por los puntos que cumplen

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} \leq 1.$$

Dicha área viene dada por

$$\int_{\mathbb{R}^2} \chi_E(x,y) \, dxdy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} \chi_E(x,y) \, dy\right) dx.$$

La función $\chi_E(x,y)$ (como función de y para un x fijo) es nula salvo si $-a \leq x \leq a$, en cuyo caso vale 0 salvo si $-b\sqrt{1-(x/a)^2} \leq y \leq b\sqrt{1-(x/a)^2}$, y en este caso vale 1. Por consiguiente la última integral es

$$\int_{-a}^{a} \left(\int_{-b\sqrt{1-(x/a)^2}}^{b\sqrt{1-(x/a)^2}} dy\right) dx = \int_{-a}^{a} 2b\sqrt{1-\left(\frac{x}{a}\right)^2} dx$$

El cambio $x/a = \operatorname{sen} t$ transforma la integral en

$$2ab\int_{-\pi/2}^{\pi/2} \cos^2 t \, dt = 2ab\int_{-\pi/2}^{\pi/2} \frac{1+\cos 2t}{2} \, dt = 2ab\left[\frac{t}{2} + \frac{\operatorname{sen} 2t}{4}\right]_{-\pi/2}^{\pi/2} = \pi ab.$$

En particular, el área de un círculo de radio r es πr^2. ■

Ejemplo Calculemos ahora el volumen de una esfera de radio r. Concretamente, sea $S = \{(x,y,z) \in \mathbb{R}^3 \mid x^2+y^2+z^2 \leq r^2\}$. Hemos de calcular

$$\int_{\mathbb{R}^3} \chi_S(x,y,z) \, dxdydz = \int_{-r}^{r} \left(\int_{\mathbb{R}^2} \chi_S(x,y,z) \, dydz\right) dx.$$

Fijado $x \in \,]-r, r[$, la función $(y, z) \mapsto \chi_S(x, y, x)$ es la función característica de un círculo de radio $\sqrt{r^2 - x^2}$ y la integral interior es el área de este círculo, o sea, vale $\pi(r^2 - x^2)$. Así pues, el volumen de la esfera es

$$\int_{-r}^{r} \pi(r^2 - x^2)\, dx = \left[\pi r^2 x - \frac{\pi x^3}{3}\right]_{-r}^{r} = \frac{4}{3}\,\pi r^3.$$

■

Ejemplo Vamos a generalizar el cálculo anterior. Sea

$$B_r^n = \{x \in \mathbb{R}^n \mid x_1^1 + \cdots + x_n^2 \leq r\}.$$

Vamos a probar que $m(B_r^n) = v_n r^n$, para una cierta constante v_n que también calcularemos. Razonaremos por inducción sobre n. El ejemplo anterior es el caso $n = 3$, para el cual $v_n = (4/3)\pi$, Claramente $v_2 = \pi$. En general, si χ_n es la función característica de B_r^n y fijamos $x_1 \in \,]-r, r[$, la función $(\chi_n)_{x_1}$ es la función característica de una bola en $\mathbb{R}^{n-1}$ de radio $\sqrt{r^2 - x_1^2}$. Por hipótesis de inducción

$$m(B_r^n) = \int_{-r}^{r} v_{n-1}(r^2 - x_1^2)^{(n-1)/2}\, dx_1.$$

Hacemos el cambio de variable $x_1 = r \operatorname{sen} \theta$, con lo que

$$m(B_r^n) = v_{n-1} r^n \int_{-\pi/2}^{\pi/2} \cos^n \theta \, d\theta.$$

Si llamamos $\kappa_n = \int_{-\pi/2}^{\pi/2} \cos^n \theta \, d\theta$, hemos probado que $m(B_r^n) = v_{n-1}\kappa_n\, r^n$, luego el resultado es cierto para n con $v_n = v_{n-1}\kappa_n$.

Con la notación del ejemplo de la página 147 tenemos que $\kappa_n = [I_{0,n}]_{-\pi/2}^{\pi/2}$. Si suponemos $n \geq 2$ los cálculos de dicho ejemplo nos dan que

$$\kappa_n = \frac{n-1}{n}\, \kappa_{n-2}.$$

Una simple inducción nos da ahora que

$$\kappa_n \kappa_{n-1} = \frac{2\pi}{n}.$$

En efecto, basta usar la relación anterior y tener en cuenta que

$$\kappa_0 = \int_{-\pi/2}^{\pi/2} d\theta = \pi, \quad \kappa_1 = \int_{-\pi/2}^{\pi/2} \cos\theta \, d\theta = 2.$$

Por consiguiente $v_n = v_{n-1}\kappa_n = v_{n-2}\kappa_n\kappa_{n-1}$. Así llegamos a las relaciones

$$v_1 = 2, \quad v_2 = \pi, \quad v_n = \frac{2\pi}{n}\, v_{n-1},$$

que nos permiten calcular fácilmente v_n para cualquier valor de n. ■

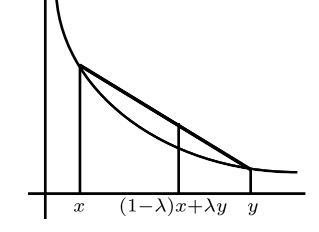
8.2 Espacios L^p

En las secciones posteriores vamos a necesitar algunos resultados abstractos referentes a ciertos espacios de funciones integrables que estudiaremos aquí. Introducimos primero una noción elemental del análisis de una variable que nos va a ser de gran ayuda:

Definición 8.11 Una función $f : I \longrightarrow \mathbb{R}$ definida en un intervalo abierto es *convexa* si cuando $x, y \in I$ y $0 < \lambda < 1$ entonces

$$f\big((1-\lambda)x + \lambda y\big) \leq (1-\lambda)f(x) + \lambda f(y).$$

La interpretación geométrica es clara: la gráfica de f ha de quedar por debajo del segmento que une los puntos $\big(x, \phi(x)\big)$ e $\big(y, \phi(y)\big)$.



Llamando $z = (1-\lambda)x + \lambda y$ y despejando λ vemos que la condición de convexidad equivale a que si $x < z < y$ entonces

$$(y-x)f(z) \leq (y-z)f(x) + (z-x)f(y),$$

lo cual a su vez equivale a

$$\frac{f(z)-f(x)}{z-x} \leq \frac{f(y)-f(z)}{y-z}.$$

Por el teorema del valor medio, de aquí se sigue que si f es derivable y f' es una función monótona creciente, entonces f es convexa. El recíproco también es cierto y fácil de probar, pero no lo necesitaremos.

Diremos que dos números reales positivos p y q son *conjugados* si

$$\frac{1}{p} + \frac{1}{q} = 1.$$

Es obvio que cada $p > 1$ tiene un único conjugado $q > 1$. El caso $p = q = 2$ es especialmente importante. Como los pares de conjugados aparecen normalmente como exponentes, es frecuente llamarlos “exponentes conjugados”.

Teorema 8.12 (Desigualdad de Hölder) *Sean p y q exponentes conjugados. Sea X un espacio medida y sean $f, g : X \longrightarrow [0, +\infty]$ funciones medibles. Entonces*

$$\int_X fg\, d\mu \leq \left(\int_X f^p\, d\mu\right)^{1/p} \left(\int_X g^q\, d\mu\right)^{1/q}.$$

Demostración: Llamemos A y B a los dos factores del segundo miembro. Si $A = 0$ entonces el teorema 7.22 implica que $f = 0$ p.c.t.p., luego $fg = 0$ p.c.t.p. y la desigualdad es clara. Si $A > 0$ y $B = +\infty$ de nuevo es obvio.

Podemos suponer, pues, $0 < A < +\infty$ y $0 < B < +\infty$. Llamemos $F = f/A$, $G = g/B$. Entonces

$$\int_X F^p \, d\mu = \int_X G^p \, d\mu = 1. \tag{8.1}$$

Supongamos que x tal que $0 < F(x) < +\infty$, $0 < G(x) < +\infty$. Entonces existen números s y t tales que $F(x) = e^{s/p}$, $G(x) = e^{t/q}$. Como $1/p + 1/q = 1$ y la función exponencial es convexa, concluimos que

$$e^{s/p+t/q} \leq p^{-1}e^s + q^{-1}e^t.$$

Por consiguiente,

$$F(x)G(x) \leq p^{-1}F(x)^p + q^{-1}G(x)^q.$$

Esta desigualdad es trivialmente cierta si $G(x) = 0$ o $G(x) = 0$, luego vale para todo x. Integrando y usando (8.1) resulta

$$\int_X FG \, d\mu \leq p^{-1} + q^{-1} = 1.$$

De aquí se sigue inmediatamente la desigualdad de Hölder. ■

Teorema 8.13 (Desigualdad de Minkowski) *Sea X un espacio medida y $f, g : X \longrightarrow [0, +\infty]$ funciones medibles. Para todo $p \geq 1$ se cumple*

$$\left(\int_X (f+g)^p \, d\mu\right)^{1/p} \leq \left(\int_X f^p \, d\mu\right)^{1/p} + \left(\int_X g^p \, d\mu\right)^{1/p}$$

Demostración: Podemos suponer que $p > 1$, que el primer miembro es mayor que 0 y que el segundo en menor que $+\infty$. Como la función x^p es convexa en $]0, +\infty[$ tenemos que

$$\left(\frac{f+g}{2}\right)^p \leq \frac{1}{2}(f^p + g^p),$$

con lo que el primer miembro también es finito, es decir, las tres integrales son finitas. Sea q el conjugado de p. Escribimos

$$(f+g)^p = f(f+g)^{p-1} + g(f+g)^{p-1}.$$

Aplicamos la desigualdad de Hölder junto con que $(p-1)q = p$ (porque p y q son conjugados). El resultado es

$$\int_X f(f+g)^{p-1} \, d\mu \leq \left(\int_X f^p \, d\mu\right)^{1/p} \left(\int_X (f+g)^p \, d\mu\right)^{1/q}.$$

Intercambiando los papeles de f y g y sumando las desigualdades resulta

$$\int_X (f+g)^p \, d\mu \leq \left(\int_X (f+g)^p \, d\mu\right)^{1/q} \left(\left(\int_X f^p \, d\mu\right)^{1/p} + \left(\int_X g^p \, d\mu\right)^{1/p}\right).$$

Dividiendo entre el primer factor del segundo miembro y teniendo en cuenta que $1 - 1/q = 1/p$ tenemos la desigualdad buscada. ■

Definición 8.14 Sea X un espacio medida, $1 \leq p < +\infty$ y $f : X \longrightarrow \mathbb{R}$ una función medible. Sea

$$\|f\|_p = \left(\int_X |f|^p \, d\mu\right)^{1/p}.$$

Llamaremos $L^p(\mu)$ al conjunto de todas las funciones medibles f tales que $\|f\|_p < +\infty$. Notemos que $L^1(\mu)$ es el conjunto de todas las funciones integrables, tal y como ya lo teníamos definido.

Por la desigualdad de Minkowski tenemos que si $f, g \in L^p(\mu)$ entonces $\|f+g\|_p \leq \|f\|_p + \|g\|_p$. En particular $f + g \in L^p(\mu)$. Por otra parte es claro que $\|\alpha f\|_p = |\alpha| \, \|f\|_p$, con lo que $\alpha f \in L^p(\mu)$. En particular vemos que $L^p(\mu)$ es un espacio vectorial sobre $\mathbb{R}$.

No es cierto que $\| \;\|_p$ sea una norma en $L^p(\mu)$, porque existen funciones no nulas f tales que $\|f\|_p = 0$ (las que son nulas p.c.t.p.). Ahora bien, es claro que las funciones de "norma" nula forman un subespacio vectorial de $L^p(\mu)$. Usaremos también la notación $L^p(\mu)$ para referirnos al espacio vectorial cociente. Si dos funciones f y g están en la misma clase entonces $f = g + h$, donde $\|h\|_p = 0$, luego $\|f\|_p \leq \|g\|_p + 0$ e igualmente tenemos la desigualdad contraria, luego $\|f\|_p = \|g\|_p$.

Podemos definir la norma de una clase de funciones como la norma de cualquiera de sus miembros. Al considerar clases de equivalencia sí tenemos un espacio normado, pues las funciones de norma 0 forman una única clase.

Teorema 8.15 *Sea X un espacio medida y $1 \leq p < +\infty$. Entonces $L^p(\mu)$ es un espacio de Banach.*

DEMOSTRACIÓN: Sea $\{f_n\}_{n=1}^{\infty}$ una sucesión de Cauchy en $L^p(\mu)$. Basta probar que tiene una subsucesión convergente. Extrayendo una subsucesión podemos suponer que $\|f_{n+1} - f_n\| < 2^{-n}$. Sea

$$g_k = \sum_{n=1}^{k} |f_{n+1} - f_n|, \quad g = \sum_{n=1}^{\infty} |f_{n+1} - f_n|.$$

Claramente $\|g_k\|_p < 1$ y aplicando el lema de Fatou a $\{g_k^p\}_{k=1}^{\infty}$ concluimos que $\|g\|_p \leq 1$. En particular $g(x) < +\infty$ p.c.t.x. Así pues, la serie

$$f(x) = f_1(x) + \sum_{n=1}^{\infty} \big(f_{n+1}(x) - f_n(x)\big)$$

converge absolutamente p.c.t.x. Definamos $f(x) = 0$ en los puntos donde no converja. Teniendo en cuenta quiénes son las sumas parciales de la serie, es claro que

$$f(x) = \lim_n f_n(x) \qquad \text{p.c.t.}x.$$

Veamos que $f \in L^p(\mu)$ y que es el límite para la norma de la sucesión dada. Dado $\epsilon > 0$ existe un k tal que si $m, n > k$ entonces $\|f_n - f_m\|_p < \epsilon$. Por el

lema de Fatou tenemos

$$\int_X |f - f_m|^p \, d\mu \leq \underline{\text{lím}}_n \int_X |f_n - f_m|^p \, d\mu \leq \epsilon^p.$$

Esto significa que $\|f - f_m\|_p \leq \epsilon$, de donde $\|f\|_p \leq \|f_k\| + \epsilon$ y por lo tanto $f \in L^p(\mu)$. También es claro ahora que f es el límite en $L^p(\mu)$ de la sucesión dada. ■

En la prueba del teorema anterior hemos visto lo siguiente:

Teorema 8.16 *Toda sucesión que converge en un espacio $L^p(\mu)$ a una función f, tiene una subsucesión que converge puntualmente a f.*

En el caso de los espacios $L^2(\mu)$ todavía podemos decir más:

Teorema 8.17 *Sea X un espacio medida. Entonces $L^2(\mu)$ es un espacio de Hilbert con el producto escalar dado por*

$$fg = \int_X fg \, d\mu.$$

DEMOSTRACIÓN: La integral que define el producto escalar es finita, pues por la desigualdad de Hölder cumple en realidad que $|fg| \leq \|f\|_2 \|g\|_2$. Claramente es bilineal y la norma que induce es precisamente la de $L^2(\mu)$. ■

Ejercicio: Sea μ la medida en $\{1, \ldots, n\}$ en la que cada punto tiene medida 1. Probar que $L^p(\mu) = \mathbb{R}^n$ y que las normas $\| \ \|_1$ y $\| \ \|_2$ son las definidas en el capítulo I.

Terminamos el estudio de los espacios $L^p(\mu)$ con dos teoremas de densidad:

Teorema 8.18 *Sea X un espacio medida. Sea S la clase de las funciones simples que son nulas salvo en un conjunto de medida finita. Entonces S es un subconjunto denso de $L^p(\mu)$ para $1 \leq p < +\infty$.*

DEMOSTRACIÓN: Es claro que $S \subset L^p(\mu)$. Tomemos primero una función $f \geq 0$ en $L^p(\mu)$. Sea $\{s_n\}_{n=1}^{\infty}$ una sucesión monótona creciente de funciones simples que converja a f. Como $0 \leq s_n \leq f$ es claro que $s_n \in L^p(\mu)$, luego $s_n \in S$. Como $|f - s_n|^p \leq f^p$, el teorema de la convergencia dominada implica que $\|f - s_n\|_p$ converge a 0, luego f está en la clausura de f. Para el caso general aplicamos la parte ya probada a f^+ y f^-. ■

Teorema 8.19 *Sea μ una medida de Borel regular en un espacio localmente compacto X. Entonces $C_c(X)$ es denso en $L^p(\mu)$ para $1 \leq p < +\infty$.*

DEMOSTRACIÓN: Consideremos la clase S del teorema anterior. Basta ver que toda función $s \in S$ puede aproximarse por una función de $C_c(X)$. Sea $\epsilon > 0$ y K el supremo de s, que claramente es finito. Por el teorema de Lusin existe una función $g \in C_c(X)$ tal que $g = s$ salvo en un conjunto de medida menor que ϵ. Por consiguiente $\|g - s\|_p \leq 2K\epsilon^{1/p}$. ■

8.3 Medidas signadas

Para enunciar más adecuadamente los próximos resultados conviene que modifiquemos nuestra definición de medida o, con más exactitud, que introduzcamos otro tipo de medidas distintas de las medidas positivas. Aunque pronto veremos la utilidad del nuevo concepto desde un punto de vista puramente matemático, quizá ahora sea más conveniente motivarlo mediante un ejemplo físico: la función que a cada región del espacio le asigna la cantidad de materia que contiene es un ejemplo de medida positiva (que podemos suponer finita), sin embargo, la aplicación que a cada región del espacio le asigna la carga eléctrica que contiene ya no se ajusta a nuestra definición de medida, porque puede tomar valores negativos, y pese a ello puede tratarse de forma muy similar.

Definición 8.20 Sea $\mathcal{A}$ una σ-álgebra en un conjunto X. Una *medida signada (finita)* en $\mathcal{A}$ es una aplicación $\mu : \mathcal{A} \longrightarrow \mathbb{R}$ tal que $\mu(\varnothing) = 0$ y si $\{E_n\}_{n=1}^{\infty}$ es una familia de conjuntos de $\mathcal{A}$ disjuntos dos a dos entonces

$$\mu\left(\bigcup_{n=1}^{\infty} E_n\right) = \sum_{n=1}^{\infty} \mu(E_n).$$

Observar que en la definición está implícita la hipótesis de que las series de medidas de conjuntos disjuntos son convergentes (en el caso de las medidas positivas donde admitíamos el valor $+\infty$ esto era evidente). Más aún, la serie ha de converger absolutamente. En efecto, la serie (finita o infinita) formada por los términos correspondientes a los conjuntos E_n con medida negativa ha de converger a la medida de su unión, y obviamente la serie de los valores absolutos converge al valor absoluto de la suma, es decir, los términos negativos convergen absolutamente. Lo mismo vale para los términos positivos, luego la serie completa también converge absolutamente.

Conviene saber que toda la teoría que vamos a exponer sobre medidas signadas se generaliza con cambios mínimos a medidas con valores complejos, pero no vamos a necesitar nada al respecto. Con esta definición, las medidas signadas sobre una σ-álgebra fija en un conjunto X forman un espacio vectorial real con las operaciones dadas por

$$(\mu + \nu)(E) = \mu(E) + \nu(E), \quad (\alpha\mu)(E) = \alpha\mu(E).$$

En particular, las medidas signadas de Borel en un espacio topológico X forman un espacio vectorial real.

Ejemplo Sea X un espacio topológico y $x \in X$. Definimos la *delta de Dirac* de soporte x como la medida de Borel dada por

$$\delta_x(E) = \begin{cases} 1 & \text{si } x \in E \\ 0 & \text{si } x \notin E \end{cases}.$$

Claramente se trata de una medida signada positiva. Si una región del espacio está ocupada por partículas puntuales en las posiciones $x_1, \ldots, x_n$ con

cargas eléctricas $q_1, \ldots, q_n$, entonces la distribución de carga viene dada por la medida signada

$$\mu = \sum_{i=1}^{n} q_i \delta_{x_i}.$$

También podemos considerar la medida positiva

$$|\mu| = \sum_{i=1}^{n} |q_i| \delta_{x_i},$$

que a cada región del espacio le asigna la cantidad total de carga que contiene, haciendo abstracción de su signo. Vamos a probar que a toda medida signada μ le podemos asignar una medida positiva $|\mu|$ con una interpretación análoga a la de este ejemplo. ■

Definición 8.21 Sea μ una medida signada en un conjunto X. Llamaremos *variación total* de μ a la aplicación definida sobre la misma σ-álgebra que μ dada por

$$|\mu|(E) = \sup \sum_{n=1}^{\infty} |\mu(E_n)|,$$

donde el supremo se toma sobre todas las particiones $\{E_n\}_{n=1}^{\infty}$ de E en conjuntos medibles disjuntos dos a dos.

Tomando la partición formada únicamente por E obtenemos la relación $|\mu(E)| \leq |\mu|(E)$. Es inmediato comprobar que la medida $|\mu|$ construida en el ejemplo anterior es la variación total de μ en el sentido de la definición anterior.

Teorema 8.22 *La variación total de una medida compleja es una medida positiva finita.*

DEMOSTRACIÓN: Obviamente $|\mu|(\varnothing) = 0$. Sea $\{E_n\}_{n=1}^{\infty}$ una partición de un conjunto medible E en conjuntos medibles disjuntos dos a dos. Sea $r_n < |\mu|(E_n)$. Entonces cada E_n tiene una partición $\{E_{nm}\}_{m=1}^{\infty}$ de modo que

$$r_n < \sum_{m=1}^{\infty} |\mu(E_{nm})|.$$

La unión de todas las particiones forma una partición de E, con lo que

$$\sum_{n=1}^{\infty} r_n \leq \sum_{m,n=1}^{\infty} |\mu(E_{nm})| \leq |\mu|(E).$$

Tomando el supremo en todas las posibles elecciones de $\{r_n\}_{n=1}^{\infty}$ resulta que

$$\sum_{n=1}^{\infty} |\mu|(E_n) \leq |\mu|(E).$$

Sea ahora $\{A_m\}_{m=1}^{\infty}$ una partición de E en conjuntos medibles disjuntos dos a dos. Entonces $\{E_n \cap A_m\}_{n=1}^{\infty}$ es una partición de A_m y , $\{E_n \cap A_m\}_{m=1}^{\infty}$ es una partición de E_n, luego

$$\begin{aligned}\sum_{m=1}^{\infty} |\mu(A_m)| &= \sum_{m=1}^{\infty} \left| \sum_{n=1}^{\infty} \mu(A_m \cap E_n) \right| \leq \sum_{m=1}^{\infty} \sum_{n=1}^{\infty} |\mu(A_m \cap E_n)| \\ &= \sum_{n=1}^{\infty} \sum_{m=1}^{\infty} |\mu(A_m \cap E_n)| \leq \sum_{n=1}^{\infty} |\mu|(E_n).\end{aligned}$$

Como esto vale para toda partición de E, tenemos

$$|\mu|(E) \leq \sum_{n=1}^{\infty} |\mu|(E_n).$$

Falta probar que $|\mu|$ es finita. Supongamos que existe un conjunto medible E tal que $|\mu|(E) = +\infty$. Sea $t = 2\big(1 + |\mu(E)|\big)$. Puesto que $|\mu|(E) > t$, por definición de variación total existen conjuntos medibles E_n contenidos en E y disjuntos dos a dos tales que

$$t < \sum_{n=1}^{k} |\mu(E_n)|.$$

Sea P la suma de los términos $|\mu(E_n)|$ tales que $\mu(E_n) \geq 0$ y sea N la suma de los $|\mu(E_n)|$ tales que $\mu(E_n) < 0$. Por la desigualdad anterior tenemos $t < P + N$, luego $t < 2P$ o bien $t < 2N$, según si $N \leq P$ o $P \leq N$. Sea A la unión de los E_n correspondientes a P o N según el caso, de modo que $A \subset E$ y $t < 2|\mu(A)|$, luego $|\mu(A)| > t/2 > 1$.

Sea ahora $B = E \setminus A$. Entonces

$$|\mu(B)| = |\mu(B) - \mu(A)| \geq |\mu(A)| - |\mu(E)| > \frac{t}{2} - |\mu(E)| = 1.$$

Así pues, hemos partido E en dos conjuntos disjuntos A y B tales que $|\mu(A)| > 1$ y $|\mu(B)| > 1$. Obviamente $|\mu|(A) = +\infty$ o bien $|\mu|(B) = +\infty$.

Supongamos ahora que el espacio total X tiene variación total infinita. Por el argumento anterior podemos partirlo en dos conjuntos medibles disjuntos $X = A_1 \cup B_1$ tales que $|\mu|(B_1) = +\infty$ y $|\mu(A_1)| > 1$. Aplicando el mismo razonamiento a B_1 obtenemos $B_1 = A_2 \cap B_2$ con $|\mu|(B_2) = +\infty$ y $|\mu(A_2)| > 1$. Procediendo de este modo construimos una familia numerable de conjuntos medibles disjuntos $\{A_n\}_{n=1}^{\infty}$ tales que $|\mu(A_n)| > 1$ para todo n. Debería cumplirse

$$\mu\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n),$$

pero la serie no converge, porque su término general no tiende a 0. Esta contradicción prueba que $|\mu|(X) < +\infty$ y por lo tanto $|\mu|$ es una medida finita. ■

Ejercicio: Probar que el conjunto de todas las medidas signadas sobre una σ-álgebra en un conjunto X es un espacio normado con la norma dada por $\|\mu\| = |\mu|(X)$.

Definición 8.23 Si μ es una medida signada en un conjunto X, llamaremos *variación positiva* y *variación negativa* de μ a las medidas (definidas sobre la misma σ-álgebra) dadas por

$$\mu^+ = \frac{|\mu| + \mu}{2}, \quad \mu^- = \frac{|\mu| - \mu}{2}.$$

Claramente son dos medidas positivas finitas y cumplen las relaciones

$$\mu = \mu^+ - \mu^-, \quad |\mu| = \mu^+ + \mu^-.$$

Por ejemplo, si μ representa la carga eléctrica contenida en una región del espacio, μ^+ y μ^- representan, respectivamente, la carga positiva y la carga negativa que contiene dicha región.

Diremos que una función $f : X \longrightarrow \mathbb{R}$ es *integrable* respecto a una medida signada μ si lo es respecto a μ^+ y μ^-, y definimos su integral como

$$\int_X f\,d\mu = \int_X f\,d\mu^+ - \int_X f\,d\mu^-.$$

Es fácil ver que el conjunto $L^1(\mu)$ de las funciones integrables es un espacio vectorial y la integral determina sobre él una aplicación lineal. Además se cumple la desigualdad

$$\left|\int_X f\,d\mu\right| \leq \int_X |f|\,d|\mu|.$$

También es claro que la aplicación dada por $\nu(E) = \int_E f\,d\mu$ es una medida signada en X.

Nos encaminamos a probar ahora uno de los teoremas más importantes de la teoría de la medida. Para ello necesitamos algunos conceptos y resultados previos.

Definición 8.24 sea μ una medida positiva en un conjunto X y λ una medida arbitraria (positiva o signada) en la misma σ-álgebra. Diremos que λ es *absolutamente continua* respecto a μ, y lo representaremos por $\lambda \ll \mu$, si todos los conjuntos nulos para μ son nulos para λ.

Si existe un conjunto medible A tal que para todo conjunto medible E se cumple $\lambda(E) = \lambda(A \cap E)$ se dice que λ *está concentrada* en A. Esto equivale a que $\lambda(E) = 0$ siempre que $E \cap A = \varnothing$.

Diremos que dos medidas arbitrarias (sobre una misma σ-álgebra) son *mutuamente singulares,* y lo representaremos por $\lambda_1 \perp \lambda_2$, si existen conjuntos medibles disjuntos A y B tales que λ_1 está concentrada en A y λ_2 está concentrada en B.

Ejemplo Si admitimos como principio que toda masa ocupa un volumen, entonces la medida μ que a cada región del espacio le asigna la masa que contiene es absolutamente continua respecto a la medida de Lebesgue m. Por el contrario, a veces es más conveniente trabajar con masas puntuales, es decir, suponiéndolas localizadas en puntos del espacio sin volumen. Éste sería el caso de una distribución de masas de la forma $\mu = \sum_{n=1}^{k} m_n \delta_{x_n}$. Es fácil ver que entonces $\mu \perp m$. ■

He aquí algunas propiedades elementales:

Teorema 8.25 *Sean λ, λ_1 y λ_2 medidas arbitrarias en un conjunto X y μ una medida positiva, todas ellas con los mismos conjuntos medibles. Entonces*

a) Si λ está concentrada en un conjunto medible A, también lo está $|\lambda|$.

b) Si $\lambda_1 \perp \lambda_2$ entonces $|\lambda_1| \perp |\lambda_2|$.

c) Si $\lambda_1 \perp \mu$ y $\lambda_2 \perp \mu$ entonces $\lambda_1 + \lambda_2 \perp \mu$.

d) Si $\lambda_1 \ll \mu$ y $\lambda_2 \ll \mu$ entonces $\lambda_1 + \lambda_2 \ll \mu$.

e) Si $\lambda \ll \mu$, entonces $|\lambda| \ll \mu$.

f) Si $\lambda_1 \ll \mu$ y $\lambda_2 \perp \mu$ entonces $\lambda_1 \perp \lambda_2$.

g) Si $\lambda \ll \mu$ y $\lambda \perp \mu$ entonces $\lambda = 0$.

DEMOSTRACIÓN: a) Si $E \cap A = \varnothing$ y $\{E_n\}_{n=1}^{\infty}$ es cualquier partición de E, entonces $\lambda(E_n) = 0$ para todo n, luego $|\lambda|(E) = 0$.

b) Es consecuencia inmediata de a).

c) Existen conjuntos disjuntos A_1 y B_1 tales que λ_1 está concentrada en A_1 y μ está concentrada en B_1 e igualmente existen conjuntos disjuntos A_2 y B_2 tales que λ_2 está concentrada en A_2 y μ está concentrada en B_2. Entonces $\lambda_1 + \lambda_2$ está concentrada en $A = A_1 \cap A_2$ y μ está concentrada en $B = B_1 \cap B_2$.

d) Obvio.

e) Si $\mu(E) = 0$ y $\{E_n\}_{n=1}^{\infty}$ es una partición de E en conjuntos medibles disjuntos, entonces $\mu(E_n) = 0$ para todo n, luego $\lambda(E_n) = 0$ y $|\lambda|(E) = 0$.

f) Tenemos que λ_2 está concentrada en un conjunto A tal que $\mu(A) = 0$. Como $\lambda_1 \ll \mu$ ha de ser $\lambda_1(E) = 0$ para todo conjunto medible $E \subset A$. Por consiguiente λ_1 está concentrada en $X \setminus A$.

g) Por f) tenemos $\lambda \perp \lambda$, pero esto implica que $\lambda = 0$. ■

Ahora probamos dos hechos elementales sobre medidas positivas que nos harán falta a continuación.

Teorema 8.26 *Si μ es una medida positiva σ-finita en un conjunto X, entonces existe una función $w \in L^1(\mu)$ tal que $0 < w(x) < 1$ para todo $x \in X$.*

Demostración: Sea $\{E_n\}_{n=1}^{\infty}$ una partición de X en conjuntos disjuntos de medida finita. Definamos

$$w_n(x) = \begin{cases} \frac{1}{2^n(1+\mu(E_n))} & \text{si } x \in E_n \\ 0 & \text{si } x \in X \setminus E_n \end{cases}$$

La función $w = \sum\limits_{n=1}^{\infty} w_n$ cumple lo pedido. ■

Teorema 8.27 *Sea μ una medida positiva finita, $f \in L^1(\mu)$ y $C \subset \mathbb{R}$ un conjunto cerrado. Si*

$$P_E(f) = \frac{1}{\mu(E)} \int_E f \, d\mu \in C$$

para todo conjunto medible E no nulo, entonces $f(x) \in C$ p.c.t. $x \in X$.

Demostración: Sea $I = [x-r, x+r]$ un intervalo cerrado disjunto de C. Puesto que $\mathbb{R} \setminus C$ es unión de una familia numerable de tales intervalos, basta probar que $E = f^{-1}[I]$ es nulo. En caso contrario

$$|P_E(f) - x| = \frac{1}{\mu(E)} \left| \int_E (f-x) \, d\mu \right| \leq \frac{1}{\mu(E)} \int_E |f-x| \, d\mu \leq r,$$

lo cual es imposible, pues $P_E(f) \in C$. ■

Finalmente podemos probar:

Teorema 8.28 (de Lebesgue-Radon-Nikodým) *Sea μ una medida positiva σ-finita en un conjunto X y sea λ una medida signada sobre la misma σ-álgebra. Entonces*

a) Existe un único par de medidas signadas λ_a y λ_s tales que

$$\lambda = \lambda_a + \lambda_s, \quad \lambda_a \ll \mu, \quad \lambda_s \perp \mu.$$

Si λ es positiva (y finita) también lo son λ_a y λ_s.

b) Existe una única $h \in L^1(\mu)$ tal que para todo conjunto medible E

$$\lambda_a(E) = \int_E h \, d\mu.$$

La parte a) se conoce como Teorema de Lebesgue. La parte b) es el Teorema de Radon-Nikodým.

Demostración: La unicidad de a) es clara a partir de 8.25, pues si λ_a', λ_s' es otro par que cumpla lo mismo, entonces $\lambda_a' - \lambda_a = \lambda_s - \lambda_s'$, $\lambda_a' - \lambda_a \ll \mu$ y $\lambda_s - \lambda_s' \perp \mu$, luego $\lambda_a' - \lambda_a = \lambda_s - \lambda_s' = 0$.

La unicidad de h en b) (como función de $L^1(\mu)$, es decir, p.c.t.p.) es fácil de probar: si existe otra h' en las mismas condiciones entonces $f = h - h'$ tiene

integral nula sobre todo conjunto. Tomamos $E = \{x \in X \mid f(x) > 0\}$ y, como f tiene integral nula sobre E, por el teorema 7.22 concluimos que $f = 0$ p.c.t.p.

Supongamos primero que λ es positiva (y finita). Sea w según el teorema 8.26. Sea ν la medida positiva finita dada por

$$\nu(E) = \lambda(E) + \int_E w\, d\mu.$$

En otros términos, si $f = \chi_E$ se cumple

$$\int_X f\, d\nu = \int_X f\, d\lambda + \int_X fw\, d\mu.$$

Claramente, la misma relación vale cuando f es una función simple y, en consecuencia, para funciones medibles no negativas. Si $f \in L^2(\nu)$ la desigualdad de Hölder implica

$$\left|\int_X f\, d\lambda\right| \leq \int_X |f|\, d\lambda \leq \int_X |f|\, d\nu \leq \left(\int_X |f|^2\, d\nu\right)^{1/2} \nu(X)^{1/2} = \nu(X)^{1/2}\|f\|_2.$$

Según el teorema 2.37, esto significa que

$$f \mapsto \int_X f\, d\lambda$$

es una aplicación lineal continua de $L^2(\nu)$ en $\mathbb{R}$. Como $L^2(\nu)$ es un espacio de Hilbert, el teorema 2.44 implica que existe $g \in L^2(\nu)$ tal que para toda $f \in L^2(\nu)$ se cumple

$$\int_X f\, d\lambda = \int_X fg\, d\nu. \tag{8.2}$$

Si aplicamos esto a una función característica $f = \chi_E$, donde $\nu(E) > 0$ el miembro izquierdo es $\lambda(E)$ y así

$$0 \leq \frac{1}{\nu(E)} \int_E g\, d\nu = \frac{\lambda(E)}{\nu(E)} \leq 1.$$

Según el teorema 8.27, resulta que $g(x) \in [0, 1]$ p.c.t.x (respecto a ν). Puesto que g sólo está determinada como elemento de $L^2(\nu)$, podemos modificarla en un conjunto nulo y suponer que $0 \leq g(x) \leq 1$ para todo $x \in X$. Entonces (8.2) puede reescribirse como

$$\int_X (1 - g)f\, d\lambda = \int_X fgw\, d\mu. \tag{8.3}$$

Sean

$$A = \{x \in X \mid 0 \leq g(x) < 1\}, \quad B = \{x \in X \mid g(x) = 1\}.$$

Definimos las medidas λ_a y λ_s mediante

$$\lambda_a(E) = \lambda(A \cap E), \quad \lambda_s(E) = \lambda(B \cap E).$$

Haciendo $f = \chi_B$ en (8.3) el miembro izquierdo es 0 y el derecho es $\int_B w\,d\mu$, y como $w > 0$ concluimos que $\mu(B) = 0$, luego $\lambda_s \perp \mu$.

Ahora aplicamos (8.3) a $(1 + g + \cdots + g^n)\chi_E$, con lo que tenemos

$$\int_E (1 - g^{n+1})\,d\lambda = \int_E g(1 + g + \cdots + g^n) w\,d\mu.$$

El integrando de la izquierda es nulo en B y converge a 1 en A, luego la integral de la izquierda converge a $\lambda(A \cap E) = \lambda_a(E)$. Por otra parte, el integrando de la derecha converge a una función medible no negativa h (quizá con valores infinitos), luego tomando límites en n resulta que

$$\lambda_a(E) = \int_E h\,d\mu.$$

En particular esto vale para $E = X$, lo que prueba que h toma valores finitos p.c.t.p., luego modificándola si es necesario en un conjunto nulo podemos suponer que $h \in L^1(\mu)$.

Esto prueba el teorema cuando λ es positiva. Si es una medida signada arbitraria basta aplicar la parte ya probada a λ^+ y λ^-. ■

Definición 8.29 Si μ una medida positiva σ-finita en un conjunto X y λ es una medida signada sobre la misma σ-álgebra tal que $\lambda \ll \mu$, entonces la función h cuya existencia afirma el teorema de Radon-Nikodým se llama la *derivada de Radon-Nikodým* de λ respecto a μ. La relación que expresa el teorema se representa también por $d\lambda = h\,d\mu$.

Veamos una aplicación del teorema de Radon-Nikodým. Una interpretación del teorema siguiente es que si μ representa la distribución de carga eléctrica en el espacio, entonces el espacio puede dividirse en dos regiones, una íntegramente ocupada por cargas positivas y otra por cargas negativas.

Teorema 8.30 (Teorema de descomposición de Hahn) *Sea μ una medida signada en un conjunto X. Entonces X se descompone en unión de dos conjuntos medibles disjuntos A y B tales que para todo conjunto medible E se cumple*

$$\mu^+(E) = \mu(A \cap E), \quad \mu^-(E) = -\mu(B \cap E).$$

DEMOSTRACIÓN: Obviamente $\mu \ll |\mu|$, luego existe una función $h \in L^1(|\mu|)$ tal que $d\mu = h\,d|\mu|$. Veamos que h toma los valores ± 1 p.c.t.p. Dado un número real r, sea $A_r = \{x \in X \mid |h(x)| < r\}$. Para toda partición $\{E_n\}_{n=1}^\infty$ de A_r se cumple

$$\sum_{n=1}^\infty |\mu(E_n)| = \sum_{n=1}^\infty \left| \int_{E_n} h\,d|\mu| \right| \leq \sum_{n=1}^\infty r|\mu|(E_n) = r|\mu|(A_r).$$

Para $r < 1$ esto implica que $|\mu|(A_r) = 0$, luego $|h| \geq 1$ p.c.t.p. Por otra parte, si $|\mu|(E) > 0$ tenemos que

$$\left| \frac{1}{|\mu|(E)} \int_E h\,d|\mu| \right| = \frac{|\mu(E)|}{|\mu|(E)} \leq 1,$$

luego el teorema 8.27 implica que $|h| \leq 1$ p.c.t.p. Modificando h en un conjunto nulo podemos suponer que $h = \pm 1$. Ahora basta definir

$$A = \{x \in X \mid h(x) = 1\}, \quad B = \{x \in X \mid h(x) = -1\}.$$

En efecto, por definición de μ^+ tenemos que para todo conjunto medible E se cumple

$$\mu^+(E) = \frac{1}{2}\int_E (1+h)\,d|\mu| = \int_{E\cap A} h\,d|\mu| = \mu(E\cap A).$$

Igualmente se razona con la variación negativa. ■

Se dice que el par (A, B) es una *partición de Hahn* de μ. Como aplicación obtenemos un par de hechos de interés sobre la derivada de Radon-Nikodým de una medida signada:

Teorema 8.31 *Sea μ una medida positiva σ-finita en un conjunto X y $f \in L^1(\mu)$. Sea λ la medida signada determinada por $d\lambda = f\,d\mu$. Entonces $|\lambda| = |f|\,d\mu$ y si $g \in L^1(\lambda)$ entonces $gf \in L^1(\mu)$ y*

$$\int_X g\,d\lambda = \int_X gf\,d\mu.$$

DEMOSTRACIÓN: Claramente $\lambda \ll \mu$, luego $|\lambda| \ll \mu$, luego $\lambda^+ \ll \mu$ y $\lambda^- \ll \mu$. Por el teorema de Radon-Nikodým existen funciones f_+, $f_- \in L^1(\mu)$ tales que $d\lambda^+ = f_+\,d\mu$, $d\lambda^- = f_-\,d\mu$. Si (A, B) es una partición de Hahn para λ, podemos exigir que f_+ se anule en B y f_+ se anule en A. Es claro que $f = f_+ - f_m$ p.c.t.p., luego $|f| = f_+ + f_-$ p.c.t.p. Por consiguiente $d|\lambda| = (f_+ + f_-)\,d\mu = |f|\,d\mu$.

La segunda parte del teorema es obvia si g es una función simple. Si g es positiva tomamos una sucesión creciente $\{s_n\}$ de funciones simples $0 \leq s_n \leq g$ que converja puntualmente a g. Entonces

$$\int_X s_n|f|\,d\mu = \int_X s_n\,d|\lambda|.$$

Por el teorema de la convergencia monótona concluimos que

$$\int_X g|f|\,d\mu = \int_X g\,d|\lambda| < +\infty,$$

luego $gf \in L^1(\mu)$. También tenemos

$$\int_X s_n\,d\lambda^+ - \int_X s_n\,d\lambda^- = \int_X s_n f\,d\mu$$

Aplicando el teorema de la convergencia monótona a la izquierda y el de la convergencia dominada a la derecha llegamos a la igualdad del enunciado. Si g no es positiva aplicamos la parte ya probada a g^+ y g^-. ■

Para terminar probaremos una versión del teorema de Riesz para medidas signadas. Sea K un espacio topológico compacto y $C(K)$ el espacio de todas las funciones reales continuas sobre K. Sabemos que $C(K)$ es un espacio de Banach con la norma supremo. Sea $C(K)'$ el espacio de las aplicaciones lineales continuas de $C(K)$ en $\mathbb{R}$, que también es un espacio de Banach con la norma dada por

$$\|T\| = \sup\big\{|T(f)| \bigm| \|f\|_\infty \leq 1\big\}.$$

Además, para toda $f \in C(K)$ se cumple $|T(f)| \leq \|T\|\,\|f\|_\infty$. Llamemos $M(K)$ al conjunto de todas las medidas signadas de Borel en K, que claramente es un espacio normado con la norma $\|\mu\| = |\mu|(K)$.

Teorema 8.32 (Teorema de representación de Riesz) *Si K es un espacio compacto, a cada funcional lineal continuo $T \in C(K)'$ le corresponde una única medida signada $\mu \in M(K)$ tal que para toda función $f \in C(K)$ se cumple*

$$T(f) = \int_K f\,d\mu.$$

Además esta correspondencia es una isometría $C(K)' \longrightarrow M(K)$.

Demostración: Si T fuera positivo, es decir, si $T(f) \geq 0$ cuando $f \geq 0$, la versión del teorema de Riesz que probamos en el capítulo anterior nos daría la medida que buscamos. En el caso general vamos a descomponer T en diferencia de dos funcionales positivos. Sea $C^-(K) = \{f \in C(K) \mid f \geq 0\}$ y definamos

$$T^+(f) = \sup\{T(u) \mid u \in C^+(K),\ u \leq f\}, \quad \text{para } f \in C^+(K).$$

Notar que si $0 \leq u \leq f$ se cumple $|T(u)| \leq \|T\|\,\|u\|_\infty \leq \|T\|\,\|f\|_\infty$, luego T^+ es finito y $|T^+(f)| \leq \|T\|\,\|f\|_\infty$.

Es claro que si $\alpha \geq 0$ se cumple $T^+(\alpha f) = \alpha T^+(f)$. Además $T^+(f+g) = T^+(f) + T^+(g)$. En efecto, si $0 \leq u \leq f$ y $0 \leq v \leq g$ entonces $0 \leq u+v \leq f+g$, luego $T(u)+T(v) \leq T^+(f+g)$. Tomando supremos $T^+(f)+T^+(g) \leq T^+(f+g)$. Recíprocamente, si $w \leq f+g$ es claro que $u = f \wedge w$ y $v = w - u$ son funciones continuas y $0 \leq u \leq f$, $0 \leq v \leq g$, $w = u+v$, luego $T(w) \leq T^+(f) + T^+(g)$ y, tomando supremos, $T^+(f+g) \leq T^+(f) + T^+(g)$.

Dada $f \in C(K)$ definimos $T^+(f) = T^+(f^+) - T^+(f^-)$. Es fácil probar que $T^+ : C(K) \longrightarrow \mathbb{R}$ es un funcional lineal continuo positivo. Lo mismo vale para $T^- = T^+ - T$. Por el teorema de Riesz para funcionales positivos existen medidas positivas μ_+ y μ_- tales que

$$T^+(f) = \int_K f\,d\mu_+, \quad T^-(f) = \int_K d\mu_-.$$

Ambas medidas son finitas (basta aplicar las fórmulas a $f = 1$). Por lo tanto podemos definir $\mu = \mu_+ - \mu_-$, que es una medida signada en K y claramente representa a T.

Veamos que $\|\mu\| = \|T\|$. Si $\|f\|_\infty \leq 1$ tenemos

$$|T(f)| = \left|\int_K f\,d\mu\right| \leq \int_K |f|\,d|\mu| \leq |\mu|(K) = \|\mu\|,$$

luego $\|T\| \leq \|\mu\|$. Dado $\epsilon > 0$ consideramos una partición de Hahn (A, B) para μ. Sean K_1 y K_2 conjuntos compactos tales que $K_1 \subset A$, $K_2 \subset B$, $\mu^+(A) - \mu^+(K_1) + \mu^-(B) - \mu^-(K_2) < \epsilon$. Es fácil construir una función continua $f : X \longrightarrow [-1, 1]$ que valga 1 sobre K_1 y -1 sobre K_2. Entonces

$$\|\mu\| = \int_A d\mu^+ + \int_B d\mu^- \leq \int_K f\,d\mu + \epsilon = T(f) + \epsilon \leq \|T\| + \epsilon,$$

luego $\|\mu\| \leq \|T\|$. En particular tenemos que la correspondencia $T \mapsto \mu$ es inyectiva (su núcleo es trivial) y obviamente es suprayectiva. ■

8.4 Derivación de medidas

Supongamos que μ representa la distribución de la masa en el espacio y m es la medida de Lebesgue. Si suponemos que $\mu \ll m$, es decir, si la materia ocupa un volumen, entonces tiene sentido hablar de la densidad de materia en un punto x del espacio, entendida como la cantidad de materia por unidad de volumen. Una aproximación a dicha densidad es el cociente

$$\frac{\mu(B)}{m(B)},$$

donde B es un entorno de x. Sin embargo, si la distribución de la materia no es uniforme, dicho cociente no es exactamente la densidad, pero se parecerá más a ella cuanto menor sea el entorno considerado. Estas ideas nos llevan a la definición siguiente:

Definición 8.33 Sea m la medida de Lebesgue en $\mathbb{R}^n$ y μ una medida de Borel signada en $\mathbb{R}^n$. Definimos los cocientes

$$C_r\mu(x) = \frac{\mu(B_r(x))}{m(B_r(x))},$$

donde las bolas las tomamos respecto a la distancia euclídea. Definimos la *derivada* de μ en x como

$$\frac{d\mu}{dm}(x) = \lim_{r \to 0} C_r\mu(x).$$

Probaremos que si $\mu \ll m$ entonces la derivada existe p.c.t.p. Para ello nos apoyaremos en la *función maximal* M, definida por

$$M\mu(x) = \sup_{0<r<+\infty} C_r|\mu|(x).$$

Veamos que es medible usando para ello el teorema 7.7. Sea E la antiimagen por $M\mu$ de un intervalo $]t, +\infty]$, es decir, $E = \{x \in X \mid M\mu(x) > t\}$. Veamos que es abierto. Dado $x \in E$, existe un $r > 0$ tal que $u = C_r|\mu|(x) > t$, luego $\mu(B_r(x)) = um(B_r(x))$. Tomemos $\delta > 0$ tal que $(r + \delta)^n < r^n u/t$. Así, si $|y - x| < \delta$ entonces $B_r(x) \subset B_{r+\delta}(y)$, con lo que

$$|\mu|(B_{r+\delta}(y)) \geq um(B_r(x)) = u\left(\frac{r}{r+\delta}\right)^n m(B_{r+\delta}(y)) > tm(B_{r+\delta}(y)).$$

Esto prueba que $y \in E$, es decir, tenemos que $B_\delta(x) \subset E$, luego E es abierto. ■

Ahora necesitamos un par de hechos técnicos:

Teorema 8.34 *Sea W la unión de una familia finita de bolas $B_{r_i}(x_i) \subset \mathbb{R}^n$, para $i = 1, \ldots, N$. Entonces existe un conjunto $S \subset \{1, \ldots, n\}$ tal que:*

a) Las bolas $B_{r_i}(x_i)$ con $i \in S$ son disjuntas,

b) $W \subset \bigcup\limits_{i\in S} B_{3r_i}(x_i)$,

c) $m(W) \leq 3^n \sum\limits_{i\in S} m(B_{r_i}(x_i))$.

Demostración: Escribiremos $B_i = B_{r_i}(x_i)$. Ordenemos las bolas de modo que sus radios sean decrecientes. Sea $i_1 = 1$. Eliminemos todas las bolas que corten a B_{i_1}. Sea B_{i_2} la primera bola restante, si es que queda alguna, eliminemos las bolas que cortan a B_{i_2} y continuemos el proceso hasta que no queden bolas. Veamos que las bolas que hemos dejado cumplen el teorema. Ciertamente son disjuntas. Cada bola B_j de las que hemos eliminado está contenida en una bola $B_{3r_i}(x_i)$, para algún $i \in S$, pues si $r' \leq r$ y $B_{r'}(x')$ corta a $B_r(x)$ entonces $B_{r'}(x') \subset B_{3r}(x)$.

La parte c) es consecuencia inmediata de b). ■

Teorema 8.35 *Si μ es una medida signada de Borel en $\mathbb{R}^n$ y $t > 0$, entonces*

$$m\big(\{x \in \mathbb{R}^n \mid M\mu(x) > t\}\big) \leq \frac{3^n}{t}\,|\mu|(\mathbb{R}^n).$$

Demostración: Sea K un subconjunto compacto del abierto que aparece en el miembro izquierdo. Cada $x \in K$ es el centro de una bola abierta B tal que $|\mu|(B) > tm(B)$.

Estas bolas forman un cubrimiento de K, del cual podemos extraer un subcubrimiento finito al que a su vez podemos aplicar el teorema anterior, digamos $\{B_1, \ldots, B_k\}$ de modo que

$$m(K) \leq 3^n \sum_{i=1}^n m(B_i) \leq \frac{3^n}{t} \sum_{i=1}^n |\mu|(B_i) \leq \frac{3^n}{t}\,|\mu|(\mathbb{R}^n).$$

La medida μ es regular porque lo son sus variaciones positiva y negativa, luego tomando el supremo sobre todos los compactos K obtenemos la desigualdad del enunciado. ■

Si $f \in L^1(\mathbb{R}^n)$ podemos aplicar el teorema anterior a la medida definida por

$$\mu(E) = \int_E |f| \, dm.$$

En este caso $M\mu$ es la función

$$Mf(x) = \sup_{0<r<+\infty} \frac{1}{m(B_r(x)} \int_{B_r(x)} |f| \, dm$$

y la tesis del teorema es

$$m\big(\{x \in \mathbb{R}^n \mid Mf(x) > t\}\big) \leq \frac{3^n}{t} \, \|f\|_1. \tag{8.4}$$

La existencia de la derivada de una medida se deducirá de un teorema de existencia de puntos de Lebesgue, que definimos a continuación:

Definición 8.36 Sea $f \in L^1(\mathbb{R}^n)$ (representaremos así al espacio $L^1(m)$, para la medida de Lebesgue en $\mathbb{R}^n$). Un *punto de Lebesgue* de f es un punto $x \in \mathbb{R}^n$ tal que

$$\lim_{r \to 0} \frac{1}{m(B_r(x))} \int_{B_r(x)} |f(y) - f(x)| \, dm(y) = 0.$$

Notemos que si x es un punto de Lebesgue entonces, dado que

$$\left| \frac{1}{m(B_r(x))} \int_{B_r(x)} f \, dm - f(x) \right| = \left| \frac{1}{m(B_r(x))} \int_{B_r(x)} \big(f(y) - f(x)\big) \, dm(y) \right|$$

$$\leq \frac{1}{m(B_r(x))} \int_{B_r(x)} |f(y) - f(x)| \, dm(y),$$

se cumple

$$f(x) = \lim_{r \to 0} \frac{1}{m(B_r(x))} \int_{B_r(x)} f \, dm.$$

Es claro que si f es continua en x entonces x es un punto de Lebesgue para f, pero necesitamos la existencia de puntos de Lebesgue de funciones integrables cualesquiera:

Teorema 8.37 *Si $f \in L^1(\mathbb{R}^n)$, entonces casi todo $x \in \mathbb{R}^n$ es un punto de Lebesgue de f.*

Demostración: Sea

$$T_r(f)(x) = \frac{1}{m(B_r(x))} \int_{B_r(x)} |f - f(x)|\, dm$$

y sea

$$T(f)(x) = \overline{\lim_{r\to 0}}\, T_r(f)(x).$$

Tenemos que probar que $Tf = 0$ p.c.t.p. Fijemos un número real $y > 0$ y un número natural k. Por el teorema 8.19 existe una función $g \in C_c(\mathbb{R}^n)$ tal que $\|f - g\| - 1 < 1/k$. Sea $h = f - g$. La continuidad de g implica que $T(g) = 0$. Como

$$T_r(h)(x) \leq \frac{1}{m(B_r(x))} \int_{B_r(x)} |h|\, dm + |h(x)|,$$

tenemos que $T(h) \leq Mh + |h|$. Por otra parte, dado que $T_r(f) \leq T_r(g) + T_r(h)$, vemos que $T(f) \leq Mh + |h|$. Así pues,

$$\{x \in \mathbb{R}^n \mid T(f)(x) > 2y\} \subset \{x \in \mathbb{R}^n \mid M(h)(x) > y\} \cup \{x \in \mathbb{R}^n \mid |h|(x) > y\}$$

Llamemos $E(y, k)$ al miembro derecho de la inclusión anterior. Por 8.4 tenemos que la medida del primero de los conjuntos de la unión es menor o igual que $3^n/(yk)$. Respecto al segundo, llamémoslo A, observamos que

$$ym(A) \leq \int_A |h|\, dm \leq \int_{R^n} |h|\, dm = \|h\|_1 < \frac{1}{k},$$

luego en total, $m\big(E(y, k)\big) \leq (3^n + 1)/(yk)$.

El conjunto $\{x \in \mathbb{R}^n \mid T(f)(x) > 2y\}$ es independiente de k y está contenido en la intersección de los conjuntos $E(y, k)$ para todo k, que es nula. Por la completitud de la medida de Lebesgue concluimos que es medible Lebesgue y tiene medida nula. Como esto vale para todo $y > 0$ concluimos que $Tf = 0$ p.c.t.p. ■

Con esto llegamos al teorema principal de esta sección:

Teorema 8.38 *Sea μ una medida signada de Borel en $\mathbb{R}^n$ tal que $\mu \ll m$ y sea f la derivada de Radon-Nikodým de μ respecto a m. Entonces $d\mu/dm = f$ p.c.t.p. y para todo conjunto de Borel $E \subset \mathbb{R}^n$ se cumple*

$$\mu(E) = \int_E \frac{d\mu}{dm}\, dm.$$

Demostración: El teorema de Radon-Nikodým afirma que se verifica la igualdad del enunciado con f en lugar de $d\mu/dm$. Para cada punto de Lebesgue x de f se cumple

$$f(x) = \lim_{r\to 0} \frac{1}{m(B_r(x))} \int_{B_r(x)} f\, dm = \lim_{r\to 0} \frac{\mu(B_r(x))}{m(B_r(x))} = \frac{d\mu}{dm}(x).$$

■

8.5 El teorema de cambio de variable

En esta sección probaremos un teorema fundamental para el cálculo de integrales, junto con el teorema de Fubini. Se trata de la generalización a funciones de varias variables de la regla de integración por sustitución. El planteamiento es el siguiente: Supongamos que $g : U \longrightarrow V$ es un difeomorfismo entre dos abiertos en $\mathbb{R}^n$, el problema es relacionar la integral de una función $f : V \longrightarrow \mathbb{R}$ con la de $g \circ f$. Según el teorema 7.35, si g fuera una aplicación lineal de determinante Δ se cumpliría que $m\big(g(A)\big) = |\Delta|\, m(A)$, para todo conjunto medible $A \subset U$. En este caso no es difícil deducir que

$$\int_V f\, dm = |\Delta| \int_U (g \circ f)\, dm.$$

En el caso general, sabemos que en un entorno de cada punto x la aplicación g se confunde con su diferencial $dg(x)$, que es una aplicación lineal de determinante $\Delta_g(x) = \det Jg(x)$ (el *determinante jacobiano* de g en x). Esto se traduce en que si A es un conjunto medible contenido en un entorno de x suficientemente pequeño, entonces $m\big(g(A)\big) \approx |\Delta_g(x)|\, m(A)$. Esto es suficiente para llegar a un resultado análogo al caso lineal:

$$\int_V f\, dm = \int_U (g \circ f)|\Delta_g|\, dm.$$

Éste es el contenido del teorema de cambio de variable. La prueba detallada no es trivial en absoluto, sino que depende de una gran parte de los resultados que hemos visto hasta ahora. Observemos que $dg(x)$ es de hecho un isomorfismo, luego $\Delta_g(x) \neq 0$. Comencemos probando la relación entre la medida de un conjunto y la de su imagen para el caso de bolas abiertas:

Teorema 8.39 *Sea $g : U \longrightarrow V$ un difeomorfismo entre dos abiertos de $\mathbb{R}^n$ y $x \in U$. Entonces*

$$\lim_{r \to 0} \frac{m\big(g[B_r(x)]\big)}{m\big(B_r(x)\big)} = |\Delta_g(x)|.$$

Demostración: Puesto que la medida de Lebesgue es invariante por traslaciones, no perdemos generalidad si suponemos $x = 0$ y $g(0) = 0$. Sea $\phi = dg(0)$ y $h = g \circ \phi^{-1}$. Probaremos que

$$\lim_{r \to 0} \frac{m\big(h[B_r(0)]\big)}{m\big(B_r(0)\big)} = 1.$$

El teorema 7.35 nos da que $m\big(h[B_r(0)]\big) = |\Delta_g(0)|^{-1} m\big(g[B_r(x)]\big)$, luego la igualdad anterior implica la que figura en el enunciado. La aplicación h cumple $h(0) = 0$ y además $dh(0)$ es la aplicación identidad. Por definición de diferenciabilidad esto significa que

$$\lim_{x \to 0} \frac{h(x) - x}{\|x\|} = 0.$$

Así, dado $0 < \epsilon < 1$ existe un $\delta > 0$ tal que

$$\text{si } \|x\| < \delta \text{ entonces } \|h(x) - x\| < \epsilon\|x\|. \tag{8.5}$$

Como h es un difeomorfismo en particular es una aplicación abierta, luego podemos tomar δ de modo que además $B_\delta(0)$ está contenida en la imagen de h. Veamos que si $0 < r < \delta$ entonces

$$B_{(1-\epsilon)r}(0) \subset h[B_r(0)] \subset B_{(1+\epsilon)r}(0). \tag{8.6}$$

En efecto, si $y \in B_{(1-\epsilon)r}(0) \subset B_\delta(0)$ entonces existe un $x \in U$ tal que $h(x) = y$. La relación 8.5 implica $\|y - x\| < \epsilon\|x\|$. Entonces

$$\|x\| \leq \|x - y\| + \|y\| < \epsilon\|x\| + (1 - \epsilon)r,$$

luego $(1 - \epsilon)\|x\| < (1 - \epsilon)r$ y $\|x\| < r$. Así $y \in h[B_r(0)]$. Por otra parte, si $y \in h[B_r(0)]$, es decir, si $y = h(x)$ con $\|x\| < r$, la relación 8.5 implica que $\|y - x\| < \epsilon r$, luego $\|y\| < (1 + \epsilon)r$, lo que nos da la otra inclusión.

Tomando medidas en 8.6 resulta

$$(1 - \epsilon)^n \leq \frac{m\big(h[B_r(0)]\big)}{m\big(B_r(0)\big)} \leq (1 + \epsilon)^n, \quad \text{para todo } r < \delta,$$

y la conclusión es clara. ■

En las condiciones del teorema anterior la aplicación g biyecta claramente los conjuntos de Borel de U con los de V, por lo que podemos definir una medida de Borel en U mediante $\mu(A) = m\big(g[A]\big)$. Si se trata de una medida finita el teorema afirma que

$$\frac{d\mu}{dm}(x) = \Delta_g(x).$$

En realidad no hay ningún problema en definir la derivada de una medida positiva, aunque no sea finita, pues se trata de un concepto local, y la igualdad anterior es cierta en cualquier caso. Necesitaremos probar que $\mu \ll m$, lo que se seguirá del teorema siguiente:

Teorema 8.40 *La imagen de un conjunto nulo por una función diferenciable es un conjunto nulo.*

Demostración: Sea $g : U \longrightarrow V$ diferenciable y sea $E \subset U$ un conjunto nulo. Si $x \in E$, la diferenciabilidad de g en x significa que para $y \neq x$

$$g(y) - g(x) = \|y - x\| \left(dg(x)\left(\frac{y - x}{\|y - x\|}\right) + E(y - x) \right)$$

donde E es una cierta función continua en 0 con $E(0) = 0$. Como $dg(x)$ está acotada en la bola unidad, existen naturales k y p tales que

$$\text{si} \quad y \in B_{1/p}(x) \quad \text{entonces} \quad \|g(y) - g(x)\| \leq k\|y - x\|.$$

Sea F_{kp} el conjunto de todos los $x \in E$ que cumplen esta relación. Hemos probado que E está contenido en la unión de estos conjuntos. Por consiguiente basta probar que $g[F_{kp}]$ es nulo.

Sea M el cociente entre la medida de una bola de radio 1 y la de un cubo de diámetro 1. Una homotecia de razón r los transforma en una bola de radio r y un cubo de diámetro r, cuyas medidas difieren de las anteriores en la constante r^n, luego la razón entre ambas sigue siendo M, es decir, M es en realidad el cociente entre la medida de una bola de radio arbitrario r y la medida de un cubo de diámetro r.

Sea $\epsilon > 0$. Cubramos F_{kp} por un abierto W tal que $m(W) < \epsilon/M$. Por el teorema 7.32 podemos descomponer W en una unión numerable de cubos disjuntos, que podemos tomar con diámetro menor que $1/p$. Desechamos los que no cortan a F_{kp} y así tenemos a éste cubierto por una familia numerable de cubos C_i cuyas medidas suman menos de ϵ/M. Cubrimos cada cubo por una bola $B_i = B_{r_i}(x_i)$ cuyo centro es un punto $x_i \in F_{kp} \cap C_i$ y su radio es el diámetro de C_i (menor que $1/p$). De este modo las bolas cubren a F_{kp} y la suma de sus medidas es menor que ϵ.

Si $x \in F_{kp} \cap B_i$, entonces $\|x - x_i\| < 1/p$ y $x_i \in F_{kp}$, luego

$$\|g(x) - g(x_i)\| \leq k\|x - x_i\| < nr_i,$$

luego $g[F_{kp} \cap B_i] \subset B_{kr_i}\big(g(x_i)\big)$. Esto prueba que $g[F_{kp}]$ está cubierto por las bolas $B_{kr_i}\big(g(x_i)\big)$, luego

$$m\big(g[F_{kp}]\big) \leq \sum_{i=1}^{\infty} m\big(B_{kr_i}\big(g(x_i)\big)\big) = k^n \sum_{i=1}^{\infty} m(B_i) < k^n\epsilon.$$

Como ϵ es arbitrario, tenemos que $m\big(g[F_{kp}]\big) = 0$. ■

Es claro que todo conjunto medible Lebesgue se puede expresar como unión de un conjunto de Borel con un conjunto nulo (el conjunto de Borel es la unión de una sucesión de compactos que aproximen la medida del conjunto dado). El teorema anterior prueba que si $g : U \longrightarrow V$ es un difeomorfismo entre dos abiertos de $\mathbb{R}^n$ y $E \subset U$ es medible Lebesgue, entonces $g[E]$ es medible Lebesgue. Veamos finalmente el teorema principal:

Teorema 8.41 (Teorema de cambio de variable) *Sea $g : U \longrightarrow V$ un difeomorfismo entre dos abiertos de $\mathbb{R}^n$ y sea $f : V \longrightarrow \mathbb{R}$ una aplicación integrable Lebesgue. Entonces*

$$\int_V f\, dm = \int_U (g \circ f)|\Delta_g|\, dm.$$

DEMOSTRACIÓN: Para cada natural k, sea $U_k = \{x \in U \mid \|g(x)\| < k\}$. Claramente U_k es abierto. Para cada conjunto medible E definimos $\mu_k(E) = m\big(g[E \cap U_k]\big)$. Claramente μ_k es una medida finita sobre la σ-álgebra de los conjuntos medibles Lebesgue en $\mathbb{R}^n$. El teorema anterior prueba que $\mu_k \ll m$.

Ahora podemos aplicar el teorema 8.38, según el cual existe $d\mu_k/dm$ p.c.t.p., es integrable Lebesgue y para todo conjunto medible E se cumple

$$\mu_k(E) = \int_E \frac{d\mu_k}{dm}\, dm.$$

En principio 8.38 prueba esto para conjuntos de Borel, pero la igualdad se extiende obviamente a conjuntos medibles arbitrarios. Del teorema 8.39 se sigue fácilmente que si $x \in U_k$ entonces

$$\frac{d\mu_k}{dm} = |\Delta_g(x)|.$$

En total hemos probado que si E es medible entonces

$$m\big(g[E \cap U_k]\big) = \int_{U_k} \chi_E |\Delta_g|\, dm.$$

Por el teorema de la convergencia monótona concluimos que

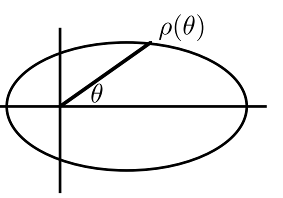
$$m\big(g[E \cap U]\big) = \int_U \chi_E |\Delta_g|\, dm. \tag{8.7}$$

Vamos a deducir de aquí que si A es un conjunto medible, entonces

$$\int_V \chi_A\, dm = \int_U (g \circ \chi_A)|\Delta_g|\, dm.$$

Basta tomar $E = g^{-1}(A) \subset U$. El comentario previo al teorema prueba que E es medible y claramente $\chi_E = g \circ \chi_A$. Además $g[E \cap U] = g[E] = A \cap V$, luego (8.7) se convierte en la igualdad buscada.

De aquí se sigue la fórmula del enunciado para el caso en que f es una función simple no negativa. Por el teorema de la convergencia monótona llegamos al mismo resultado para funciones medibles no negativas y a su vez se extiende a toda función integrable aplicándolo a f^+ y f^-. ■



Ejemplo Consideremos una curva cerrada en $\mathbb{R}^2$ que rodee a $(0,0)$ y admita una expresión en coordenadas polares $\rho = \rho(\theta)$ (es decir, que corte a cada semirrecta de origen $(0,0)$ en un único punto). El recinto S limitado por la curva estará formado por los puntos (ρ_0, θ_0) tales que $0 \leq \rho_0 \leq \rho(\theta_0)$. Para calcular el área de S efectuamos el cambio de variables $x = \rho\cos\theta$, $y = \rho \operatorname{sen}\theta$, cuyo jacobiano es ρ. Así, el área se puede calcular como

$$\int_S dxdy = \int_0^{2\pi}\int_0^{\rho(\theta)} \rho\, d\rho d\theta = \int_0^{2\pi} \frac{\rho^2}{2}\, d\theta.$$

(Hemos aplicado el teorema anterior en el abierto

$$\{(\rho, \theta) \mid 0 < \rho < \rho(\theta), 0 < \theta < 2\pi\}.$$

El cambio de coordenadas lo transforma biyectivamente en S menos el radio $\theta = 0$, que tiene área nula, por lo que no importa despreciarlo.)

Por ejemplo, el área de la cardioide $\rho = (a/2)(1 + \cos\theta)$ viene dada por

$$\frac{a^2}{8}\int_0^{2\pi}(1+\cos\theta)^2 d\theta = \frac{a^2}{8}\int_0^{2\pi}\left(1 + 2\cos\theta + \frac{1+\cos 2\theta}{2}\right)d\theta$$

$$= \frac{a^2}{8}\left[\theta + 2\operatorname{sen}\theta + \frac{\theta}{2} + \frac{\operatorname{sen} 2\theta}{4}\right]_0^{2\pi} = \frac{3\pi}{8}a^2.$$

■

Ejemplo En el capítulo VI demostramos que los cuerpos sometidos a la acción gravitatoria de una estrella o planeta describen trayectorias rectas o cónicas, pero no calculamos la posición del cuerpo en función del tiempo. Ahora probaremos la segunda ley de Kepler, que aporta información a este respecto. Se refiere a un cuerpo (un planeta, un cometa) que describe una trayectoria cónica alrededor (digamos) del Sol:

> *El radio que une el móvil con el Sol barre áreas iguales en tiempos iguales.*

Tomemos como origen la posición del Sol y sea $\rho(\theta)$ la trayectoria del móvil. Sea A el sector de cónica que barre el radio que une al móvil con el Sol entre un ángulo θ_0 y un ángulo θ_1. El área de A es

$$\int_A dxdy = \frac{1}{2}\int_{\theta_0}^{\theta_1}\rho^2\, d\theta.$$

Hacemos el cambio $\theta = \theta(t)$, donde t es el tiempo. El resultado es

$$\frac{1}{2}\int_{t_0}^{t_1}\rho^2(\theta(t))\theta'(t)\, dt = \frac{1}{2}\int_{t_0}^{t_1}\rho^2(t)\omega(t)\, dt = \frac{1}{2m}\int_{t_0}^{t_1} L\, dt = \frac{L}{2m}(t_1 - t_0).$$

Así pues, el área barrida es proporcional al tiempo recorrido.

A su vez de aquí se deduce la tercera ley de Kepler, válida para móviles que describen órbitas elípticas alrededor de un mismo cuerpo. El período de revolución de tal cuerpo es el tiempo que tarda en recorrer una órbita completa:

> *Los cuadrados de los períodos de revolución son proporcionales a los cubos de los semiejes de las órbitas.*

Según vimos en el capítulo VI, la ecuación de la órbita es

$$\rho = \frac{L^2}{GMm^2}\,\frac{1}{1+\epsilon\cos\theta}.$$

Los vértices mayores (los valores máximo y mínimo de ρ) se corresponden con los ángulos $\theta = 0, \pi$. Su semisuma es el semieje mayor:

$$a = \frac{L^2}{GMm^2}\frac{1}{2}\left(\frac{1}{1+\epsilon} - \frac{1}{1-\epsilon}\right) = \frac{L^2}{GMm^2}\frac{1}{1-\epsilon^2}.$$

Puesto que ϵ es la excentricidad, el semieje menor es $b = a\sqrt{1-\epsilon^2}$. El área de la elipse es

$$A = \pi ab = \pi a^2\sqrt{1-\epsilon^2} = \frac{\pi}{G^2M^2}\frac{L^4}{m^4(1-\epsilon^2)^2}\sqrt{1-\epsilon^2}.$$

Según hemos calculado, el período T cumple $A = LT/(2m)$, luego

$$T = \frac{2\pi}{G^2M^2}\frac{L^3}{m^3}\frac{\sqrt{1-\epsilon^2}}{(1-\epsilon^2)^2}.$$

Reuniendo todo esto vemos que

$$\frac{T^2}{a^3} = \frac{4\pi^2}{GM},$$

luego tenemos la proporción buscada. ■

Ejemplo Vamos a probar que

$$\int_{-\infty}^{+\infty} e^{-x^2}\,dx = \sqrt{\pi}.$$

Aparentemente se trata de un problema de análisis de una variable real, pero el cálculo es mucho más simple si nos apoyamos en una función de dos variables. Concretamente consideramos $f(x,y) = e^{-x^2-y^2}$. Calculamos la integral de esta función en la bola de centro 0 y radio r mediante el cambio a coordenadas polares:

$$\int_{B_r(0)} e^{-x^2-y^2}\,dxdy = \int_0^{2\pi}\int_0^{\infty} e^{-\rho^2}\rho\,d\rho d\theta = 2\pi\left[-\frac{e^{-\rho^2}}{2}\right]_0^r = \pi(1-e^{-r^2}).$$

El teorema de la convergencia monótona implica que f es integrable en $\mathbb{R}^2$ y además

$$\int_{\mathbb{R}^2} e^{-x^2-y^2}\,dxdy = \pi.$$

Por otro lado podemos aplicar el teorema de Fubini, que nos da

$$\left(\int_{-\infty}^{\infty} e^{-x^2}\,dx\right)^2 = \pi,$$

luego tenemos la igualdad que buscábamos. De aquí se deducen varias integrales de interés. En primer lugar

$$\int_0^{+\infty} e^{-x^2}\, dx = \frac{\sqrt{\pi}}{2},$$

y haciendo el cambio $x = \sqrt{t}$ resulta

$$\Pi(-1/2) = \int_0^{+\infty} t^{-1/2} e^{-t}\, dt = \sqrt{\pi}.$$

Por la ecuación funcional de la función factorial concluimos que

$$\Pi(1/2) = \frac{1}{2}\,\Pi(-1/2) = \frac{\sqrt{\pi}}{2}.$$

■

Ejemplo Con el cálculo que acabamos de hacer podemos dar una expresión explícita para las constantes v_n que calculamos en en ejemplo de la página 292. En efecto, ahora es inmediato que

$$v_n = \frac{\pi^{n/2}}{\Pi(n/2)}, \tag{8.8}$$

pues esta función coincide con v_n para $n = 1$ y $n = 2$ y satisface la misma relación recurrente que v_n. En el capítulo siguiente daremos una prueba más elegante de esta fórmula. ■

Capítulo IX

Formas diferenciales

Después de haber dedicado el capítulo anterior a la integral de Lebesgue y los teoremas fundamentales de la teoría de la medida abstracta, nos ocuparemos aquí de sus aplicaciones al análisis real en la línea de los capítulos previos. Los únicos conjuntos medibles que nos van a aparecer serán obviamente de Borel: abiertos, cerrados, un círculo menos un punto o menos un radio, rectángulos semiabiertos como $]0, r] \times]0, 2\pi[$, etc. Toda función continua f en un conjunto de Borel B (extendida como 0 fuera del mismo) es claramente medible y, si f está acotada y B tiene medida finita, entonces f es integrable. Este criterio bastará en la mayoría de los casos. Nuestro primer objetivo será definir una medida sobre las variedades diferenciables que se corresponda con el concepto de longitud, área y volumen cuando la dimensión sea 1, 2 y 3 respectivamente. De este modo podremos calcular, por ejemplo, el área de una esfera o el área de un círculo en el plano proyectivo. Notemos que la longitud de una curva ya está definida. La definición general que daremos aquí coincidirá con la que ya conocemos para las variedades de dimensión 1.

9.1 Integración en variedades

Si queremos calcular el área de una esfera, lo primero que debemos preguntarnos es ¿qué es el área de una esfera? La pregunta es menos trivial de lo que parece, pues podemos calcular el área de cualquier figura plana razonable (es decir, medible Lebesgue), pero cualquier fragmento de esfera, por pequeño que sea, no es plano, y no es evidente cómo puede compararse su área con la de ninguna figura plana. Obviamente podemos definir aplicaciones de conjuntos planos en la esfera (difeomorfismos incluso, es decir, cartas de la esfera), pero nada nos garantiza que conserven el área, y difícilmente podríamos probar que así es sin tener de hecho una definición de área.

Para hacernos una idea de lo que vamos a hacer pensemos en una esfera del tamaño de la Tierra. Si situamos sobre ella una baldosa plana de un metro cuadrado, no descansaría perfectamente sobre el suelo esférico, pero, suponiendo que se apoyara en su centro, cada esquina se levantaría tan sólo 0.04 micras del

suelo. Resulta, pues, razonable considerar que el fragmento de la Tierra cubierto por la baldosa (aunque sea imperfectamente en teoría) tiene una superficie de 1m^2, salvo un error muy pequeño. Si cubrimos toda la Tierra con baldosas de 1m^2, aunque sin duda no encajarán a la perfección, el número de baldosas empleadas será una buena aproximación de la superficie de la Tierra, y el mínimo error cometido se podrá reducir arbitrariamente a base de considerar baldosas más y más pequeñas.

Para formalizar (y generalizar) esta idea empezamos observando que si E^n es un espacio vectorial euclídeo de dimensión n (por ejemplo un subespacio de $\mathbb{R}^m$ de dimensión n con el producto escalar inducido desde $\mathbb{R}^m$), entonces existe una isometría $\phi : E^n \longrightarrow \mathbb{R}^n$ que es, de hecho, un homeomorfismo entre las topologías euclídeas. Definimos los conjuntos medibles de E^n como las antiimágenes por ϕ de los conjuntos medibles Lebesgue de $\mathbb{R}^n$ y la *medida de Lebesgue* en E^n como la dada por $m(G) = m\big(\phi(G)\big)$.

Teniendo en cuenta que las isometrías en $\mathbb{R}^n$ conservan la medida de Lebesgue (tienen determinante ± 1), es inmediato comprobar que m así definida es una medida en E^n que no depende de la elección de ϕ. Claramente se corresponde con la noción de longitud, área, volumen, etc. en E^n. El teorema 7.35 puede enunciarse ahora en este contexto general:

Teorema 9.1 *Sea $\phi : E^n \longrightarrow F^n$ una aplicación lineal entre dos espacios vectoriales euclídeos de dimensión n y sea Δ_ϕ el determinante de la matriz de ϕ respecto a dos bases ortonormales cualesquiera. Entonces, para todo conjunto medible $A \subset E^n$ se cumple $m\big(\phi(A)\big) = |\Delta_\phi|\, m(A)$.*

DEMOSTRACIÓN: Sean $f : \mathbb{R}^n \longrightarrow E^n$ y $g : \mathbb{R}^n \longrightarrow F^n$ dos isometrías. Sea $h = f \circ \phi \circ g{-}1$. Claramente el determinante de h es igual a Δ_ϕ y $\phi = f^{-1} \circ h \circ g$. Basta aplicar el teorema 7.35 y el hecho de que f y g conservan la medida. ■

Así pues, si $S \subset \mathbb{R}^m$ es una variedad diferenciable de dimensión n, para cada punto $p \in S$ tenemos definida la medida de Lebesgue en el espacio tangente T_pS que es, según sabemos, una formalización adecuada del concepto de longitud, área, volumen, o una generalización natural de éste, según la dimensión n.

Consideremos una carta $X : U \longrightarrow \mathbb{R}^m$, llamemos $V = X[U]$, fijemos un punto $p = X(x)$ y consideremos la diferencial $dX(x) : \mathbb{R}^n \longrightarrow T_pS$. Para cada punto $t \in U$ tenemos que

$$X(t) \approx p + dX(x)(t - p),$$

de modo que, si t está suficientemente próximo a p, el punto $X(t) \in V$ "se confunde" con $p + dX(x)(t - p) \in p + T_pS$. Por lo tanto, si $B \subset V$ es un conjunto de Borel en un entorno suficientemente pequeño de p y $B_X = X^{-1}[B]$, el conjunto de Borel $B^t = dX(x)[B_X - x] \subset T_pS$ cumple que $p + B^t$ es prácticamente indistinguible de B.

Particularizando esto al ejemplo de la Tierra que hemos considerado antes, si B es la superficie cubierta por una baldosa que se apoya en el punto p,

entonces $p + B^t$ es la baldosa tangente a la superficie terrestre.[1] Por ello, lo que vamos a hacer es probar que existe una medida de Borel regular μ en V con la propiedad de que si $K \subset V$ es un subconjunto compacto y lo cubrimos por un conjunto finito de conjuntos de Borel B_i disjuntos dos a dos, entonces la suma de las medidas $m(B_i^t)$ se aproxima a $\mu(K)$, y la aproximación es mejor cuanto menores son los diámetros de los conjuntos B_i. Esta propiedad justifica que consideremos a $\mu(K)$ como el volumen de K.

Empezamos calculando las medidas $m(B^t)$:

Teorema 9.2 *Sea $X : U \longrightarrow \mathbb{R}^m$ una carta de una variedad S. Sea $X(x) = p$. Para cada conjunto de Borel $B \subset U$ se cumple $m\big(dX(x)[B]\big) = \Delta_X(x)\, m(B)$, donde $\Delta_X = \sqrt{\det(g_{ij})}$ y las funciones g_{ij} son los coeficientes del tensor métrico de S.*

DEMOSTRACIÓN: Sea $\phi : \mathbb{R}^m \longrightarrow \mathbb{R}^m$ una isometría que transforme $T_p(S)$ en $\mathbb{R}^n \times \{0\}$ y sea $p_n : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ la proyección en las primeras componentes. De este modo la restricción a $T_p(S)$ de $\phi \circ p_n$ es una isometría del plano tangente en $\mathbb{R}^n$, luego $m\big(dX(x)[B]\big) = m\big((dX(x) \circ \phi \circ p_n)[B]\big)$.

Sean A y P_n las matrices de ϕ y p_n en las bases canónicas. Entonces la matriz de $(dX(x) \circ \phi \circ p_n)$ es $J_X(x)AP_n$ y según el teorema 9.1 tenemos que $m\big(dX(x)[B_r]\big) = \Delta_X(x)\, m(B_r)$, donde $\Delta_X(x) = |\det(J_X(x)AP_n)|$. Ahora usamos que para toda matriz cuadrada M se cumple $|\det M| = \sqrt{\det(MM^t)}$, con lo que

$$\Delta_X(x) = \sqrt{\det(J_X(x)AP_nP_n^tA^tJ_X(x)^t)}.$$

Pero $J_X(x)AP_nP_n^tA^tJ_X(x)^t = J_X(x)AA^tJ_X(x)^t$. En efecto, el elemento (i,j) de esta matriz es el producto de $e_iJ_X(x)AP_n$ por $(e_jJ_X(x)AP_n)^t$, donde e_i y e_j son los vectores de la base canónica de $\mathbb{R}^n$, pero $e_iJ_X(x) \in T_p(S)$, luego $e_iJ_X(x)A \in \mathbb{R}^n \times \{0\}$, e igualmente $e_jJ_X(x)A \in \mathbb{R}^n \times \{0\}$, luego el producto es el mismo aunque suprimamos P_n. Como A es la matriz de una isometría, se cumple $AA^t = I$, luego concluimos que

$$\Delta_X(x) = \sqrt{\det(J_X(x)J_X(x)^t)} = \sqrt{\det(g_{ij}(x))}.$$

■

Así, con la notación previa al teorema, tenemos que $m(B^t) = \Delta_X(x)\, m(B_X)$. Ahora vamos a probar que la medida en V dada por

$$\mu(B) = \int_{X^{-1}[B]} \Delta_X \, dm$$

cumple la propiedad requerida.

[1] Al menos si elegimos la carta X de modo que la aplicación $t \mapsto p + dX(x)(t - p)$ sea la proyección ortogonal en el plano tangente. Puede probarse que podemos elegir así la carta, pero no será necesario, porque la medida que vamos a construir sobre S no dependerá de la elección de ninguna carta de S.

Claramente μ es una medida de Borel regular en V. Es fácil comprobar que si f es una función integrable en V, entonces

$$\int_V f\,d\mu = \int_U (X\circ f)\Delta_X\,dm$$

(se prueba primero para funciones simples y luego para no negativas).

Sea ahora $K\subset V$ un conjunto compacto, de modo que $K_X\subset U$ es también compacto. Dado $\epsilon>0$, como $X^{-1}\circ\Delta_X$ es uniformemente continua sobre K, existe un $\delta>0$ tal que si $p=X(x)$, $q=X(y)\in K$ cumplen $\|p-q\|<\delta$ entonces

$$|\Delta_X(x)-\Delta_X(y)|<\epsilon_0=\frac{\epsilon}{1+m(K_X)}.$$

Ahora consideramos cualquier cubrimiento finito de K por conjuntos de Borel B_i disjuntos dos a dos de diámetro menor que δ (es fácil probar que existen tales cubrimientos) y tomemos puntos $x_i\in X^{-1}[B_i]$. Así, si $x\in X^{-1}[B_i]$ tenemos que $|\Delta_X(x)-\Delta_X(x_i)|<\epsilon_0$. Por consiguiente:

$$|\sum_i m(B_i^t)-\mu(K)| = |\sum_i (m(B_i^t)-\mu(B_i))|$$

$$= \left|\sum_i (\Delta_X(x_i)\,m(X^{-1}[B_i]) - \textstyle\int_{X^{-1}[B_i]}\Delta_X\,dm)\right|$$

$$= \left|\sum_i \textstyle\int_{X^{-1}[B_i]}(\Delta_X(x_i)-\Delta_X)\,dm\right| \leq \sum_i \textstyle\int_{X^{-1}[B_i]}|\Delta_X(x_i)-\Delta_X|\,dm$$

$$\leq \epsilon_0\sum_i m(X^{-1}[B_i]) = \epsilon_0\,m(K_X)<\epsilon,$$

como había que probar.

En principio tenemos definida una medida sobre la imagen de cada carta de la variedad S, pero a continuación probamos que todas estas medidas se extienden a una única medida en S que, en particular, no depende de ninguna carta en concreto:

Teorema 9.3 *Sea $S\subset\mathbb{R}^m$ una variedad diferenciable de dimensión n. Entonces existe una única medida de Borel regular m en S tal que si $X:U\longrightarrow V$ es una carta y f es una función integrable en V, se tiene*

$$\int_V f\,dm = \int_U (X\circ f)\Delta_X\,dm, \quad \textit{donde } \Delta_X=\sqrt{\det(g_{ij})}.$$

Demostración: Consideremos dos cartas $X:U_1\longrightarrow V_1$ e $Y:U_2\longrightarrow V_2$ y sea $B\subset V_1\cap V_2$ un conjunto de Borel. Vamos a probar que

$$\int_{X^{-1}[B]}\Delta_X\,dm = \int_{Y^{-1}[B]}\Delta_Y\,dm.$$

Restringiendo las cartas podemos suponer que $V_1=V_2$, y entonces basta aplicar el teorema de cambio de variable a la función $f=X\circ Y^{-1}$. En efecto,

se cumple que $f : U_1 \longrightarrow U_2$ es un difeomorfismo y, como $f \circ Y = X$, tenemos $J_f(x)J_Y(f(x)) = J_X(x)$, luego

$$\begin{aligned} \Delta_X^2(x) &= \det(J_X(x)J_X(x)^t) = \det\big(J_f(x)J_Y(f(x))J_Y(f(x))^t J_f(x)^t\big) \\ &= \det(J_f(x))^2 \Delta_Y^2(f(x)) \end{aligned}$$

y por consiguiente $\Delta_X(x) = \Delta_Y\big(f(x)\big)|\det J_f(x)| = \Delta_Y\big(f(x)\big)|\Delta_f(x)|$. Así:

$$\int_{X^{-1}[B]} \Delta_X\, dm = \int_{U_1} \chi_{X^{-1}[B]}\Delta_X\, dm = \int_{U_1} (f \circ \chi_{Y^{-1}})(f \circ \Delta_Y)|\Delta_f|\, dm$$

$$= \int_{U_2} \chi_{Y^{-1}[B]}\Delta_Y\, dm = \int_{Y^{-1}[B]} \Delta_Y\, dm.$$

Con esto hemos probado que las distintas medidas que tenemos definidas sobre las imágenes de las cartas de S coinciden en su dominio común. El paso siguiente es “pegar” las medidas correspondientes a un número finito de cartas. Aunque intuitivamente es obvio que esto puede hacerse, formalmente conviene simplificar las comprobaciones usando el teorema de Riesz.

Supongamos que $X_i : U_i \longrightarrow V_i$, para $i = 1, \ldots, k$ son cartas de S con imágenes acotadas. Por el teorema 7.27 existe una partición de la unidad subordinada a los abiertos V_i, es decir, una familia de funciones $h_i \prec V_i$ tales que $h_1 + \cdots + h_k = 1$. Sea $V = V_1 \cup \cdots \cup V_k$. Para cada $f \in C_c(V)$ definimos

$$T(f) = \int_{V_1} h_1 f\, d\mu_1 + \cdots + \int_{V_k} h_k f\, d\mu_k,$$

donde μ_i es la medida asociada a la carta X_i. Claramente T es un operador lineal y positivo, luego existe una medida μ en V tal que para toda $f \in C_c(V)$ se cumple

$$\int_V f\, d\mu = \int_{V_1} h_1 f\, d\mu_1 + \cdots + \int_{V_k} h_k f\, d\mu_k.$$

Si en particular tomamos $f \in C_c(V_i)$ entonces $h_j f \in C_c(V_i \cap V_j)$, luego

$$\int_{V_j} h_j f\, d\mu_j = \int_{V_i \cap V_j} h_j f\, d\mu_j = \int_{V_i \cap V_j} h_j f\, d\mu_i = \int_{V_i} h_j f\, d\mu_i,$$

luego

$$\int_V f\, d\mu = \int_{V_i} (h_1 + \cdots + h_k) f\, d\mu_i = \int_{V_i} f\, d\mu_i.$$

Por la unicidad del teorema de Riesz esto prueba que la restricción de μ a V_i es precisamente μ_i, y es claro que esta propiedad determina a μ. En particular la construcción de μ no depende de la partición de la unidad escogida.

Finalmente “pegamos” todas las medidas asociadas a todas las cartas en una única medida en S. Para ello definimos un operador $T : C_c(S) \longrightarrow \mathbb{R}$. Para cada $f \in C_c(S)$ tomamos un número finito de cartas con imagen acotada cuya

unión cubra el soporte de f. Sea V la unión de las imágenes y μ_V la medida sobre U que acabamos de construir. Definimos

$$T(f) = \int_V f\,d\mu_V.$$

Es claro que $T(f)$ no depende de las cartas con que cubrimos el soporte, pues si realizamos dos cubrimientos distintos $V = V_1 \cup \cdots \cup V_k$ y $V' = V_1' \cup \cdots \cup V_k'$, entonces cada abierto $V_i \cap V_j'$ es la imagen de dos cartas que inducen la misma medida y $T(f)$ coincide con la integral de f en $V \cap V'$ respecto a la única medida que extiende a todas ellas. Teniendo esto en cuenta es fácil probar que T es lineal y positivo, con lo que existe una única medida de Borel regular m en S tal que

$$\int_S f\,dm = T(f).$$

Es claro que m extiende a la medida inducida por cualquier carta. El resto del teorema es ya inmediato. ■

La compleción de la medida construida en el teorema anterior se llama a veces *medida de Lebesgue* en la variedad S. Observemos que si tomamos $S = \mathbb{R}^n$ con la carta identidad, la medida del teorema es precisamente la medida de Lebesgue en $\mathbb{R}^n$ (pues $\Delta_X = 1$). Lo mismo es válido si S es un subespacio vectorial de $\mathbb{R}^m$ de dimensión n.

En definitiva, hemos probado que si $K \subset S$ está contenido en la imagen de una carta X y cubrimos K por un número finito de conjuntos de Borel B_i disjuntos dos a dos, las sumas $\sum\limits_i m(B_i^t)$, que en principio dependen de la carta X y de los puntos $p_i \in B_i$ respecto a los que se calculan las proyecciones, convergen cuando los diámetros de los conjuntos B_i tienden a 0, a la medida de Lebesgue $m(K)$ que no depende de la carta ni de las elecciones de los puntos p_i. Podríamos generalizar esta propiedad para compactos no contenidos en el rango de una carta y para subconjuntos de Borel arbitrarios, pero no es necesario, puesto que el "volumen" que pretendemos definir sobre la variedad S debe extender a las medidas que hemos construido sobre las imágenes de las cartas (ya que cumplen la propiedad de aproximación que hemos tomado como condición necesaria para que la definición sea aceptable) concluimos que la única medida consistente con dicha condición es la medida de Lebesgue que acabamos de definir, luego es la única definición posible de volumen en una variedad.

Demostraremos únicamente la siguiente relación local entre la medida de Lebesgue y las medidas de las proyecciones en los espacios tangentes:

Teorema 9.4 *Sea $S \subset \mathbb{R}^m$ una variedad diferenciable de dimensión n y sea $X : U \longrightarrow S \cap V$ una carta alrededor de un punto $p \in S$. Sea $x \in U$ tal que $X(x) = p$. Para cada conjunto de Borel $B \subset S \cap V$ sea $B^t = dX(x)[X^{-1}[B]]$. Entonces*

$$\lim_{B \to p} \frac{m(B)}{m(B^t)} = 1,$$

donde el límite ha de entenderse como sigue: Para todo $\epsilon > 0$ existe un entorno G de p en $S \cap V$ tal que si $B \subset G$ es un conjunto de Borel no nulo en S entonces $|m(B)/m(B^t) - 1| < \epsilon$.

Demostración: Dado $\epsilon > 0$, sea $\delta = (\epsilon/2)\Delta_X(x)$. Por la continuidad de X^{-1} y Δ_X en x existe un entorno G de p tal que si $y \in X^{-1}[G]$ entonces $|\Delta_X(y) - \Delta_X(x)| < \delta$. Si E es un conjunto de Borel no nulo contenido en G y $B_X = X^{-1}[B]$ tenemos que

$$m(B_X)(\Delta_X(x) - \delta) \leq m(B) = \int_{B_X} \Delta_X \, dm \leq m(B_X)(\Delta_X(x) + \delta),$$

luego

$$\left| \frac{m(B)}{m(B_X)} - \Delta_X(x) \right| \leq \delta < \Delta_X(x)\epsilon.$$

Por consiguiente:

$$\left| \frac{m(B)}{\Delta_X(x) m(B_X)} - 1 \right| < \epsilon,$$

pero por 9.2 tenemos que $m(B^t) = m(dX(x)[B_X]) = \Delta_X(x)m(B_X)$, con lo que

$$\left| \frac{m(B)}{m(B^t)} - 1 \right| < \epsilon.$$

■

Ejemplo Si $\alpha :]a, b[\longrightarrow \mathbb{R}^n$ es una curva parametrizada regular que no se corta a sí misma (de modo que su imagen es una variedad S de dimensión 1 con carta α) entonces $J_\alpha(t) = \alpha'(t)$, luego $\Delta_\alpha(t) = \|\alpha'(t)\|$ y por consiguiente

$$m(S) = \int_a^b \|\alpha'(t)\| \, dt$$

es la longitud de α tal y como la teníamos definida.

Si S es una superficie en $\mathbb{R}^3$ entonces el elemento de superficie se suele representar por $d\sigma = \sqrt{EG - F^2}\, dm$. Por consiguiente el área de una región C de S cubierta por la carta puede calcularse como

$$A = \int_{X^{-1}(C)} \sqrt{EG - F^2} \, dudv.$$

■

Ejemplo Vamos a calcular el área la superficie de revolución determinada por

$$X = \big(r(u)\cos v, r(u) \operatorname{sen} v, z(u)\big).$$

Sabemos que

$$E = r'(u)^2 + z'(u)^2, \quad F = 0, \quad G = r(u)^2.$$

Por lo tanto

$$A = \int_0^{2\pi} \int_{u_0}^{u_1} r(u)\sqrt{r'(u)^2 + z'(u)^2}\, dudv = 2\pi \int_{u_0}^{u_1} r(u)\sqrt{r'(u)^2 + z'(u)^2}\, du.$$

Si en particular $z(u) = u$, la fórmula se reduce a

$$A = 2\pi \int_{u_0}^{u_1} r(u)|r'(u)|\, du.$$

Por ejemplo, el área de la esfera

$$g(\phi, \theta) = (R \operatorname{sen} \phi \cos \theta, R \operatorname{sen} \phi \operatorname{sen} \theta, R \cos \phi), \quad \phi \in \left]0, \pi\right[, \theta \in \left]0, 2\pi\right[.$$

es

$$A = 2\pi \int_0^{\pi} R^2 \operatorname{sen} \phi \, d\phi = 4\pi R^2.$$

Más detalladamente, si hacemos $\rho = R\phi$ entonces ρ es la distancia del punto (ρ, θ) al polo norte y las coordenadas

$$x = R \operatorname{sen} \frac{\rho}{R} \cos \theta, \quad y = R \operatorname{sen} \frac{\rho}{R} \operatorname{sen} \theta, \quad z = R \cos \frac{\rho}{R},$$

son el análogo esférico a las coordenadas polares en el plano. El área de un círculo esférico de radio (esférico) r es

$$A_r = 2\pi \int_0^r R^2 \operatorname{sen} \frac{\rho}{R}\, d\rho = 2\pi R^2 \left(1 - \cos \frac{r}{R}\right) = 4\pi R^2 \operatorname{sen}^2 \frac{r}{2R}.$$

Así, si $r = \pi R$ recuperamos el área de la esfera $4\pi R^2$, si r es pequeño con respecto a R entonces $\operatorname{sen}(r/R) \approx r/R$ y por consiguiente $A_r \approx \pi r^2$, el área del círculo plano del mismo radio. ■

***Ejemplo** El ejemplo anterior particularizado a $R = 1$ nos da que el área de un círculo elíptico de radio r es $4\pi \operatorname{sen}^2(r/2)$. En particular el área del plano elíptico completo es 2π. Es fácil ver que el área de un bilátero de ángulo α es 2α. Dado un triángulo elíptico T de lados a, b, c y ángulos α, β, γ, el plano elíptico es la unión de T y sus tres triángulos adyacentes. Si llamamos T_a al triángulo adyacente por el lado a, tenemos que $T \cup T_a$ forman un bilátero de ángulo α, luego

$$m(T) + m(T_a) = 2\alpha.$$

Sumando las ecuaciones análogas para los otros dos triángulos adyacentes llegamos a que

$$3m(T) + m(T_a) + m(T_b) + m(T_c) = 2\alpha + 2\beta + 2\gamma.$$

Pero $m(T)+m(T_a)+m(T_b)+m(T_c) = 2\pi$, pues es el área del plano completo, y por consiguiente

$$m(T) = \alpha + \beta + \gamma - \pi.$$

Esto nos da una demostración analítica de que la suma de los ángulos de un triángulo elíptico es siempre mayor que π. ■

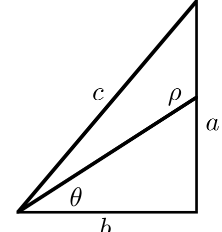
***Ejemplo** Consideremos ahora el plano hiperbólico. Según (4.6), el elemento de longitud hiperbólica en coordenadas polares es

$$ds^2 = d\rho^2 + \operatorname{senh}^2 \rho \, d\theta^2,$$

luego el elemento de área es $d\sigma = \operatorname{senh} \rho$. El área de un círculo hiperbólico de radio r es

$$2\pi \int_0^r \operatorname{senh} \rho \, d\rho = 2\pi(\cosh r - 1) = 4\pi \operatorname{senh}^2 \frac{r}{2}.$$

Consideremos ahora un triángulo rectángulo como indica la figura. Tomemos como origen de las coordenadas polares el vértice A. Para hallar su área hemos de integrar $d\sigma$ variando θ entre 0 y α y, para un θ dado, en virtud de la relación trigonométrica



$$\tanh b = \cos \theta \, \tanh \rho.$$

vemos que ρ ha de llegar hasta

$$\operatorname{argtanh} \frac{\tanh b}{\cos \theta}.$$

Por lo tanto el área es

$$\int_0^\alpha \int_0^{\operatorname{argtanh}(\tanh b/\cos\theta)} \operatorname{senh} \rho \, d\rho d\theta = \int_0^\alpha \left(\frac{1}{\sqrt{1 - \frac{\tanh^2 b}{\cos^2 \theta}}} - 1 \right) d\theta$$

$$= \int_0^\alpha \left(\frac{\cos\theta}{\sqrt{(\cos^2 \theta - 1) + (1 - \tanh^2 b)}} - 1 \right) d\theta$$

$$\int_0^\alpha \left(\frac{\cosh b \cos\theta}{\sqrt{1 - (\cosh b \operatorname{sen} \theta)^2}} - 1 \right) d\theta = \operatorname{arcsen}(\cosh b \, \operatorname{sen} \alpha) - \alpha$$

$$= \operatorname{arcsen} \cos \beta - \alpha = \frac{\pi}{2} - \beta - \alpha = \pi - \alpha - \beta - \gamma.$$

Todo triángulo hiperbólico T se puede expresar como la unión o la diferencia de dos triángulos rectángulos, de donde es fácil concluir que en general

$$m(T) = \pi - \alpha - \beta - \gamma.$$

Esto prueba que la suma de los ángulos de un triángulo hiperbólico es siempre menor que π. ■

El teorema siguiente nos conecta el teorema de Fubini con la integración en variedades:

Teorema 9.5 *Si $S_1 \subset \mathbb{R}^{m_1}$ y $S_2 \subset \mathbb{R}^{m_2}$ son variedades diferenciables, entonces la medida de Lebesgue en $S_1 \times S_2$ (restringida a los conjuntos de Borel) es el producto de las medidas de Lebesgue de S_1 y S_2 (sobre los conjuntos de Borel).*

Demostración: Sean m, m_1 y m_2 las medidas de Lebesgue en $S_1 \times S_2$, S_1 y S_2 respectivamente. Basta probar que si A_1 y A_2 son conjuntos de Borel en S_1 y S_2 entonces $m(A_1 \times A_2) = m_1(A_1)m_2(A_2)$. Es fácil ver que A_1 y A_2 se descomponen en una unión numerable disjunta de conjuntos de Borel, cada uno de los cuales está contenido en el rango de una carta. También es claro que si probamos la igualdad anterior para los productos de estos abiertos de ahí se sigue el caso general. En definitiva, podemos suponer que A_1 está contenido en el rango de una carta X_1 y A_2 está contenido en el rango de una carta X_2. Una simple comprobación nos da que $\Delta_{X_1 \times X_2} = \Delta_{X_1}\Delta_{X_2}$, luego aplicando el teorema de Fubini concluimos que

$$m(A_1 \times A_2) = \int_{X_1^{-1}[A_1]\times X_2^{-1}[A_2]} \Delta_{X_1}\Delta_{X_2}\, dx_1 \cdots dx_{n_1+n_2}$$

$$= \left(\int_{X_1^{-1}[A_1]} \Delta_{X_1}\, dx_1 \cdots dx_{n_1}\right)\left(\int_{X_2^{-1}[A_2]} \Delta_{X_2}\, dx_{n_1+1} \cdots dx_{n_1+n_2}\right)$$

$$= m_1(A_1)m_2(A_2).$$

■

Terminamos la sección con una interpretación de la curvatura de Gauss de una superficie. De hecho se trata de la definición de curvatura que adoptó el propio Gauss.

Teorema 9.6 *Sea S una superficie y p un punto en el que la curvatura no sea nula. Sea n una determinación del vector normal en un entorno de p. Entonces*

$$|K(p)| = \lim_{E \to p} \frac{m(n[E])}{m(E)},$$

donde el límite se entiende en el mismo sentido que en el teorema 9.4.

Demostración: Por el teorema 5.29 sabemos que $K(p)$ es el determinante de $dn(p)$, luego ésta diferencial es un isomorfismo. Sea g una carta alrededor de p y sea $h = g \circ n$. Es fácil ver que h puede restringirse hasta una carta alrededor de $n(p)$ en la esfera unidad. Del teorema 9.4 se sigue que

$$\lim_{E \to p} \frac{m(E)}{m(E_t)} = 1, \qquad \lim_{E \to p} \frac{m(n[E])}{m(n[E]_t)} = 1,$$

y por otra parte $n[E]_t = dn(p)[E_t]$, luego $m(n[E]_t) = |K(p)|m(E_t)$, de donde se sigue claramente el teorema ■

9.2 El álgebra exterior

Si comparamos el teorema de cambio de variable tal y como lo enunciamos en el capítulo anterior con la fórmula de una variable

$$\int_a^b u(x)\, dx = \int_{t(a)}^{t(b)} u(x(t))\, x'(t)\, dt$$

observamos una diferencia: la derivada $x'(t)$ es el determinante jacobiano de la transformación $x = x(t)$, pero aparece sin valor absoluto, con lo que un integrando positivo puede transformarse en un integrando negativo. Esto sucede cuando la función $x(t)$ es decreciente, pero entonces el intervalo $[a, b]$ se transforma en el intervalo $[t(b), t(a)]$ y, como los límites de integración aparecen invertidos, la integral se interpreta como cambiada de signo, lo cual compensa la ausencia del valor absoluto.

En el caso general también es posible eliminar el valor absoluto en el determinante jacobiano que aparece en la fórmula de cambio de variable, a condición de considerar orientados los dominios de integración (exactamente igual que un intervalo $[a, b]$ está orientado positivamente cuando $a < b$ y negativamente cuando $b < a$, y en este caso la integral se considera cambiada de signo). La ventaja de este otro enfoque es que permite emerger a una potente teoría algebraica que subyace en el cálculo integral.

Para separar esta parte puramente algebraica conviene trabajar en un espacio vectorial arbitrario E de dimensión n, tal y como hicimos al estudiar el tensor métrico de una variedad en el capítulo V. Al igual que allí, en la práctica sólo nos interesará el caso en que E es el espacio tangente $T_p(S)$ de una variedad S en un punto p. Mantendremos la misma notación que usábamos entonces: los vectores $(v_1, \ldots, v_n)$ representarán una base arbitraria de E y $(dx_1, \ldots, dx_n)$ representará su base dual, es decir, la base de E^* determinada por

$$dx_i(v_j) = \delta_{ij}.$$

En la práctica, cuando $E = T_p(S)$ la base considerada será siempre la asociada a una carta X de S alrededor de p, es decir, la formada por las derivadas parciales $D_1X(x), \ldots, D_nX(x)$, donde $X(x) = p$, con lo que su base dual será la formada por las diferenciales $dx_1(p), \ldots, dx_n(p)$, donde $x_1, \ldots, x_n$ son las funciones en S que a cada punto le asignan sus coordenadas respecto a X.

Definición 9.7 Sea E un espacio vectorial de dimensión n. Llamaremos *paralelepípedos orientados* de E a las n-tuplas $F = (v_1, \ldots, v_n) \in E^n$. A cada F le asociamos el *paralelepípedo no orientado* $P(F)$ dado por

$$P(F) = \{\alpha_1 v_1 + \cdots + \alpha_n v_n \mid \alpha_1, \ldots, \alpha_n \in [0, 1]\} \subset E.$$

Si los vectores de F son linealmente dependientes diremos que el paralelepípedo es *degenerado.*

Habiendo fijado una base en E (y considerando positiva a su orientación), podemos dividir los paralelepípedos no degenerados en positiva y negativamente orientados, según que sus vectores determinen una base con la orientación del espacio o con la contraria (es decir, según si la matriz de cambio de base respecto a la base prefijada tenga determinante positivo o negativo).

Si E es un espacio euclídeo, es claro que la medida de Lebesgue de $P(F)$ se puede calcular como el valor absoluto del determinante de las coordenadas de los vectores de F en cualquier base ortonormal B de E (pues estas coordenadas son sus imágenes a través de la isometría de E en $\mathbb{R}^n$ que transforma B

en la base canónica). Si suprimimos este valor absoluto tenemos una "medida orientada", que ya no es función de $P(F)$, sino del paralelepípedo orientado F, y además depende de la base ortonormal B seleccionada. En efecto, la medida orientada es igual a la medida de $P(F)$ si los vectores de F forman una base de E con la misma orientación que B y es dicha medida cambiada de signo en caso contrario. Si elegimos B con la misma orientación que la base $(v_1, \ldots, v_n)$ prefijada, entonces podemos considerar que el signo de la medida orientada depende de esta base y, por consiguiente, de la orientación que le hemos dado a E. El inconveniente de que la medida dependa de la orientación se ve compensado con creces por el hecho de que la medida orientada sea esencialmente un determinante y, por lo tanto, una forma multilineal alternada de E^n en $\mathbb{R}$. Las formas multilineales alternadas en un espacio euclídeo resultan ser el equivalente a las diferenciales de funciones en el caso de una variable.

Definición 9.8 Sea E un espacio vectorial euclídeo de dimensión n. Llamaremos *k-formas diferenciales (constantes)* de E (o formas de grado k) a las aplicaciones multilineales alternadas $\omega : E^k \longrightarrow \mathbb{R}$.

Multilineal quiere decir que

$$\omega(u_1, \ldots, \alpha u_i + \beta u_i', \ldots, u_k) = \alpha\omega(u_1, \ldots, u_i, \ldots, u_k) + \beta\omega(u_1, \ldots, u_i', \ldots, u_k)$$

para todo $i = 1, \ldots, n$ y alternada quiere decir que al permutar dos vectores el valor de la forma cambia de signo o, más en general, que si σ es una permutación de $\{1, \ldots, k\}$, se cumple

$$\omega(u_{\sigma(1)}, \ldots, u_{\sigma(k)}) = \operatorname{sig}\sigma\, \omega(u_1, \ldots, u_k),$$

donde $\operatorname{sig}\sigma$ es la signatura de la permutación.

Llamaremos $A^k(E)$ al conjunto de todas las k-formas de E. Claramente forman un espacio vectorial con la suma y el producto definidos puntualmente. Convendremos en que $A^0(E) = \mathbb{R}$. Definimos el *álgebra exterior* de E como la suma directa

$$A(E) = \bigoplus_{k=0}^{\infty} A^k(E).$$

Es claro que una forma alternada se anula cuando dos de sus argumentos son iguales, luego también se anula al actuar sobre vectores linealmente dependientes (al desarrollar uno como combinación lineal de los demás la imagen de la forma se descompone en una combinación lineal de imágenes de k-tuplas con dos componentes iguales). Esto implica que $A^k(E) = 0$ para $k > n$ y por lo tanto $A(E)$ tiene dimensión finita. Vamos a estudiar ahora la estructura de los espacios $A^k(E)$ para $k \leq n$.

Es claro que $A^1(E)$ es simplemente el espacio de aplicaciones lineales de E en $\mathbb{R}$, es decir, el espacio dual de E, y tiene dimensión n. Una base la forman las diferenciales $(du_1, \ldots, du_n)$.

Para obtener bases de los espacios de k-formas de orden superior introducimos el *producto exterior* de formas, definido como sigue: si $\omega \in A^k(E)$ y $\omega' \in A^{k'}(E)$, entonces $\omega \wedge \omega'$ es la $(k+k')$-forma dada por

$$(\omega\wedge\omega')(u_1,\ldots,u_{k+k'}) = \sum_{\sigma\in\Sigma_{k+k'}} \frac{\operatorname{sig}\sigma}{k!k'!}\,\omega(u_{\sigma(1)},\ldots,u_{\sigma(k)})\omega'(u_{\sigma(k+1)},\ldots,u_{\sigma(k+k')}),$$

donde σ recorre las permutaciones de $\{1,\ldots,k+k'\}$.

Es inmediato comprobar que $\omega\wedge\omega'$ es realmente una forma. Obviamente es multilineal y si $\tau\in\Sigma_{k+k'}$ entonces

$$\begin{aligned}
&(\omega\wedge\omega')(u_{\tau(1)},\ldots,u_{\tau(k+k')})\\
&\quad= \sum_{\sigma\in\Sigma_{k+k'}} \frac{\operatorname{sig}\sigma}{k!k'!}\,\omega(u_{\sigma\tau(1)},\ldots,u_{\sigma\tau(k)})\omega'(u_{\sigma\tau(k+1)},\ldots,u_{\sigma\tau(k+k')})\\
&\quad= \operatorname{sig}\tau \sum_{\sigma\in\Sigma_{k+k'}} \frac{\operatorname{sig}\sigma\tau}{k!k'!}\,\omega(u_{\sigma\tau(1)},\ldots,u_{\sigma\tau(k)})\omega'(u_{\sigma\tau(k+1)},\ldots,u_{\sigma\tau(k+k')})\\
&\quad= \operatorname{sig}\tau\,(\omega\wedge\omega')(u_1,\ldots,u_{k+k'}).
\end{aligned}$$

La definición de $\omega\wedge\omega'$ vale incluso si $k=0$ o $k'=0$. Por ejemplo, si $k=0$ convenimos que $\omega(\,)=\omega\in\mathbb{R}$, con lo que

$$(\omega\wedge\omega')(u_1,\ldots,u_{k'}) = \sum_{\sigma\in\Sigma_{k'}} \frac{\operatorname{sig}\sigma}{k'!}\,\omega\omega'(u_{\sigma(1)},\ldots,u_{\sigma(k')})$$

$$= \sum_{\sigma\in\Sigma_{k'}} \frac{1}{k'!}\,\omega\omega'(u_1,\ldots,u_{k'}) = \omega\omega'(u_1,\ldots,u_n).$$

Así pues, en este caso $\omega\wedge\omega'=\omega\omega'$ (e igualmente si $k'=0$).

Teorema 9.9 *El producto exterior tiene las propiedades siguientes (se entiende que ω, ω', ω'' son formas de los grados adecuados para que tengan sentido las operaciones):*

a) $(\omega\wedge\omega')\wedge\omega'' = \omega\wedge(\omega'\wedge\omega'')$,

b) $\omega\wedge(\omega'+\omega'') = \omega\wedge\omega'+\omega\wedge\omega''$, $(\omega+\omega')\wedge\omega'' = \omega\wedge\omega''+\omega'\wedge\omega''$,

c) $\alpha(\omega\wedge\omega') = (\alpha\omega)\wedge\omega' = \omega\wedge(\alpha\omega')$, *para* $\alpha\in\mathbb{R}$,

d) $\omega\wedge\omega' = (-1)^{kk'}\omega'\wedge\omega$.

Demostración: a) Supongamos que ω, ω' y ω'' tienen grados k, k' y k'' respectivamente. Entonces

$$\big((\omega\wedge\omega')\wedge\omega''\big)(u_1,\ldots,u_{k+k'+k''}) =$$

$$\sum_{\sigma\in\Sigma_{k+k'+k''}} \frac{\operatorname{sig}\sigma}{(k+k')!k''!}(\omega\wedge\omega')(u_{\sigma(1)},\ldots,u_{\sigma(k+k')})\omega''(u_{\sigma(k+k'+1)},\ldots,u_{\sigma(k+k'+k'')})$$
$$= \sum_{\sigma\in\Sigma_{k+k'+k''}} \frac{\operatorname{sig}\sigma}{(k+k')!k''!} \sum_{\tau\in\Sigma_{k+k'}} \frac{\operatorname{sig}\tau}{k!k'!}\,\omega(u_{\sigma\tau(1)},\ldots,u_{\sigma\tau(k)})$$
$$\omega'(u_{\sigma\tau(k+1)},\ldots,u_{\sigma\tau(k+k')})\omega''(u_{\sigma(k+k'+1)},\ldots,u_{\sigma(k+k'+k'')}).$$

Si identificamos las permutaciones $\tau \in \Sigma_{k+k'}$ con las permutaciones de $\Sigma_{k+k'+k''}$ que fijan a los índices mayores que $k+k'$ la signatura es la misma y podemos escribir $\sigma\tau$ en todos los subíndices. Entonces queda

$$\big((\omega\wedge\omega')\wedge\omega''\big)(u_1,\ldots,u_{k+k'+k''})$$
$$= \sum_{\sigma\in\Sigma_{k+k'+k''}} \sum_{\tau\in\Sigma_{k+k'}} \frac{\operatorname{sig}\sigma\tau}{(k+k')!k!k'!k''!}\,\omega(u_{\sigma\tau(1)},\ldots,u_{\sigma\tau(k)})$$
$$\omega'(u_{\sigma\tau(k+1)},\ldots,u_{\sigma\tau(k+k')})\omega''(u_{\sigma(k+k'+1)},\ldots,u_{\sigma(k+k'+k'')}).$$

Claramente, $\sigma\tau$ recorre $(k+k')!$ veces cada permutación de $\Sigma_{k+k'+k''}$, luego tenemos que

$$\big((\omega\wedge\omega')\wedge\omega''\big)(u_1,\ldots,u_{k+k'+k''})$$
$$= \sum_{\sigma\in\Sigma_{k+k'+k''}} \frac{\operatorname{sig}\sigma}{k!k'!k''!}\,\omega(u_{\sigma(1)},\ldots,u_{\sigma(k)})\omega'(u_{\sigma(k+1)},\ldots,u_{\sigma(k+k')})$$
$$\omega''(u_{\sigma(k+k'+1)},\ldots,u_{\sigma(k+k'+k'')}).$$

Si partimos de $\omega\wedge(\omega'\wedge\omega'')$ llegamos claramente a la misma expresión, luego $(\omega\wedge\omega')\wedge\omega'' = \omega\wedge(\omega'\wedge\omega'')$.

Las propiedades b) y c) son inmediatas. Para probar d) supongamos que $\omega \in A^k(E)$ y $\omega' \in A^{k'}(E)$. Entonces

$$\omega\wedge\omega' = (-1)^{kk'}\omega'\wedge\omega.$$

En efecto, sea $\tau \in \Sigma_{k+k'}$ la permutación dada por

$$\tau(i) = \begin{cases} i+k' & \text{si } i \leq k \\ i-k & \text{si } i > k \end{cases}$$

Es fácil comprobar que $\operatorname{sig}\tau = (-1)^{kk'}$. Entonces

$$(\omega\wedge\omega')(u_1,\ldots,u_{k+k'}) = \sum_{\sigma\in\Sigma_{k+k'}} \frac{\operatorname{sig}\sigma}{k!k'!}\,\omega(u_{\sigma(1)},\ldots,u_{\sigma(k)})\omega'(u_{\sigma(k+1)},\ldots,u_{\sigma(k+k')})$$
$$= \sum_{\sigma\in\Sigma_{k+k'}} \frac{\operatorname{sig}\sigma\tau}{k!k'!}\,\omega(u_{\sigma\tau(1)},\ldots,u_{\sigma\tau(k)})\omega'(u_{\sigma\tau(k+1)},\ldots,u_{\sigma\tau(k+k')})$$
$$= (-1)^{kk'} \sum_{\sigma\in\Sigma_{k+k'}} \frac{\operatorname{sig}\sigma}{k!k'!}\,\omega(u_{\sigma(k'+1)},\ldots,u_{\sigma(k'+k)})\omega'(u_{\sigma(1)},\ldots,u_{\sigma(k')})$$
$$= (-1)^{kk'}(\omega'\wedge\omega)(u_1,\ldots,u_{k+k'}).$$

■

Ahora extendemos el producto exterior hasta un producto en $A(E)$ mediante

$$\left(\sum_{i=0}^{n} \omega_i\right) \wedge \left(\sum_{j=0}^{n} \omega'_j\right) = \sum_{i,j=0}^{n} \omega_i \wedge \omega'_j.$$

Usando el teorema anterior se prueba sin dificultad que $A(E)$ es un anillo (no conmutativo) con la suma y el producto exterior. La propiedad c) vale para elementos arbitrarios de $A(E)$ (no necesariamente formas), lo que significa que el álgebra exterior es ciertamente un álgebra no conmutativa.

Podemos comparar la construcción de $A(E)$ y su estructura de álgebra con la construcción de los anillos de polinomios: las definiciones son las técnicamente necesarias para obtener las propiedades deseadas, pero una vez comprobados los hechos básicos estas definiciones pueden ser olvidadas y sustituidas por propiedades algebraicas naturales. Por ejemplo, un polinomio termina siendo una combinación de sumas y productos de indeterminadas sujetas a las propiedades de los anillos; igualmente un elemento de $A(E)$ no es más que una combinación de sumas y productos de diferenciales sujetas a las propiedades de un álgebra no conmutativa. Para ver que esto es así hemos de observar que el argumento con el que hemos probado la asociatividad del producto exterior se generaliza sin dificultad para dar una expresión simétrica del producto de un número arbitrario de formas. Sólo nos interesa el caso en que los factores son 1-formas, que queda como sigue:

Teorema 9.10 *Si $\omega_1, \ldots, \omega_k \in A^1(E)$, entonces su producto exterior es la k-forma dada por*

$$(\omega_1 \wedge \cdots \wedge \omega_k)(u_1, \ldots, u_k) = \sum_{\sigma \in \Sigma_k} \operatorname{sig} \sigma \, \omega_1(u_{\sigma(1)}) \cdots \omega_k(u_{\sigma(k)}) = \det\bigl(\omega_i(u_j)\bigr).$$

En particular

$$(dx_{i_1} \wedge \cdots \wedge dx_{i_k})(u_1, \ldots, u_k) = \det\bigl(dx_{i_r}(u_j)\bigr).$$

Más concretamente, si $i_1 < \cdots < i_k$, $j_1 < \cdots < j_k$, entonces

$$(dx_{i_1} \wedge \cdots \wedge dx_{i_k})(v_{j_1}, \ldots, v_{j_k}) = \begin{cases} 1 & \text{si } i_1 = j_1, \ldots, i_k = j_k \\ 0 & \text{en otro caso,} \end{cases}$$

pues $dx_{i_r}(v_{j_s}) = \delta_{i_r j_s}$ y el determinante será nulo en cuanto algún i_r no figure entre los j_s o viceversa. Con esto ya es fácil obtener la expresión general de una k-forma:

Teorema 9.11 *Toda forma $\omega \in A^k(E)$ se expresa de forma única como*

$$\omega = \sum_{i_1 < \cdots < i_k} \alpha_{i_1 \cdots i_k} dx_{i_1} \wedge \cdots \wedge dx_{i_k}, \quad \textit{con } \alpha_{i_1 \cdots i_k} \in \mathbb{R}.$$

Concretamente $\alpha_{i_1 \cdots i_k} = \omega(v_{i_1}, \ldots, v_{i_k}) \in \mathbb{R}$.

DEMOSTRACIÓN: Llamemos ω' al miembro derecho de la igualdad. Para probar que $\omega = \omega'$, por la multilinealidad es suficiente comprobar que ambas coinciden sobre los vectores básicos $v_{j_1}, \ldots, v_{j_k}$, y como son alternadas podemos suponer además que $j_1 < \cdots < j_k$. La observación anterior prueba que $\omega'(v_{j_1}, \ldots, v_{j_k}) = \alpha_{j_1 \cdots j_k} = \omega(v_{j_1}, \ldots, v_{j_k})$. La unicidad es clara. ■

Por consiguiente, la dimensión de $A^k(E)$ es $\binom{n}{k}$ y la dimensión del álgebra exterior $A(E)$ es 2^n. Como última observación general para operar en $A(E)$ observemos que

$$dx_i \wedge dx_j = -dx_j \wedge dx_i,$$

luego en particular $dx_i \wedge dx_i = 0$.

Ejemplo Tomemos $E = \mathbb{R}^3$ con la base canónica. Entonces una base de $A^1(\mathbb{R}^n)$ la forman las tres diferenciales dx, dy, dz, y una 2-forma arbitraria es

$$\omega = a\, dx \wedge dy + b\, dx \wedge dz + c\, dy \wedge dz, \quad \text{con } a, b, c \in \mathbb{R}.$$

Calculemos por ejemplo:

$$\begin{aligned} \omega \wedge (dx + dy) &= a\, dx \wedge dy \wedge dx + b\, dx \wedge dz \wedge dx + c\, dy \wedge dz \wedge dx \\ &+ a\, dx \wedge dy \wedge dy + b\, dx \wedge dz \wedge dy + c\, dy \wedge dz \wedge dy \\ &= c\, dx \wedge dy \wedge dz - b\, dx \wedge dy \wedge dz = (c - b)\, dx \wedge dy \wedge dz. \end{aligned}$$

Vemos así que las formas se manipulan fácilmente sin necesidad de recurrir en ningún momento a las definiciones de las operaciones, sino tan sólo usando sus propiedades algebraicas. ■

Para terminar con las generalidades sobre formas diferenciales demostraremos el teorema siguiente:

Teorema 9.12 *Sea E un espacio vectorial de dimensión n y sean $(v_1, \ldots, v_n)$, $(v'_1, \ldots, v'_n)$ dos bases de E. Sean $(dx_1, \ldots, dx_n)$ y $(dx'_1, \ldots, dx'_n)$ sus bases duales respectivas. Sea A la matriz de cambio de base (es decir, la matriz cuyas filas son las coordenadas de los vectores v_i en la segunda base). Entonces*

$$dx'_1 \wedge \cdots \wedge dx'_n = \det A\, dx_1 \wedge \cdots \wedge dx_n.$$

DEMOSTRACIÓN: Sea $A = (a_{ij})$. Entonces $v_i = a_{i1}v'_1 + \cdots + a_{in}v'_n$, de donde $dx'_j(v_i) = a_{ij}$. Por el teorema 9.10 tenemos que

$$(dx'_1 \wedge \cdots \wedge dx'_n)(v_1, \ldots, v_n) = \det\big(dx'_j(v_i)\big) = \det A\, (dx_1 \wedge \cdots \wedge dx_n)(v_1, \ldots, v_n).$$

Esto implica que ambas formas son iguales. ■

9.3 El álgebra de Grassmann

En la sección anterior hemos estudiado las formas diferenciales en un espacio vectorial. Ahora pasamos a definir formas sobre variedades diferenciables. La relación entre unas y otras es la misma que la que hay entre vectores y campos vectoriales, sólo que a los "campos de formas" se les llama simplemente "formas":

Definición 9.13 Sea S una variedad diferenciable. Una *k-forma diferencial* en S es una aplicación ω que a cada $p \in S$ le asigna una k-forma $\omega(p) \in A^k\big(T_p(S)\big)$.

Observar que una 0-forma es simplemente una función $f : S \longrightarrow \mathbb{R}$.

Si $X : U \longrightarrow V \subset S$ es una carta de S y $p \in V$, entonces una base de $A^k\big(T_p(S)\big)$ está formada por las formas

$$dx_{i_1}(p) \wedge \cdots \wedge dx_{i_k}(p), \quad \text{con } 1 \leq i_1 < \cdots < i_k \leq n.$$

Por consiguiente, para todo $p \in V$ se cumplirá que

$$\omega(p) = \sum_{1 \leq i_1 < \cdots < i_k \leq n} \alpha_{i_1 \cdots i_k}(p)\, dx_{i_1}(p) \wedge \cdots \wedge dx_{i_k}(p), \tag{9.1}$$

para ciertas funciones $\alpha_{i_1 \cdots i_k} : V \longrightarrow \mathbb{R}$, llamadas *coeficientes* de la forma en la carta dada. Si consideramos otra carta alrededor de p con coordenadas $y_1, \ldots, y_n$, entonces podremos expresar

$$dx_i(p) = \frac{\partial x_i}{\partial y_1}(p)\, dy_1(p) + \cdots + \frac{\partial x_i}{\partial y_n}(p)\, dy_n(p).$$

Al sustituir en el miembro derecho de (9.1) y desarrollar los productos exteriores obtenemos que los coeficientes $\beta_{i_1 \cdots i_k}$ de ω en la carta Y se obtienen a partir de los coeficientes en X mediante sumas y productos por las derivadas parciales que aparecen en la igualdad anterior. De aquí se sigue que las funciones $\alpha_{i_1 \cdots i_k}$ son continuas, diferenciables o de clase C^q en un punto dado si y sólo si lo son las funciones $\beta_{i_1 \cdots i_k}$ (suponiendo que las cartas sean suficientemente derivables).

Una forma diferencial de una variedad S es *continua, diferenciable* o *de clase* C^q si y sólo si lo son sus coeficientes en todas las cartas. Por razones de simplicidad en lo sucesivo sobreentenderemos que todas las variedades que consideremos tendrán cartas de clase C^∞ y por "forma diferencial" entenderemos "forma diferencial de clase C^∞". En realidad todos los resultados que probemos valdrán igualmente sin más que suponer que las cartas y las funciones consideradas son suficientemente derivables, pero no entraremos en detalles al respecto.

Si S es una variedad diferenciable, llamaremos $\Lambda^k(S)$ al conjunto de todas las formas diferenciales (de clase C^∞) en S. Es claro que se trata de un espacio vectorial con las operaciones definidas puntualmente. Definimos el *álgebra de Grassmann* de S como la suma directa

$$\Lambda(S) = \bigoplus_{k=0}^{\infty} \Lambda^k(S).$$

El producto exterior de las álgebras $A\big(T_p(S)\big)$ induce puntualmente un producto exterior en el álgebra de Grassmann, con el cual adquiere estructura de álgebra no conmutativa. Todas las propiedades del producto exterior que vimos en la sección anterior para álgebras exteriores valen trivialmente para álgebras de Grassmann.

Tenemos que $\Lambda^0(S)$ es el conjunto de las funciones de clase C^∞ definidas sobre S. Si $f \in \Lambda^0(S)$ y $\omega \in \Lambda(S)$ escribiremos $f\omega$ en lugar de $f \wedge \omega$. Notemos que $(f\omega)(p) = f(p)\omega(p)$.

Con la ayuda de estos conceptos podemos formular con precisión algunas de las ideas que exponíamos al comienzo de la sección anterior. Sea E un espacio vectorial euclídeo. Observemos que la n-forma $dx_1 \wedge \cdots \wedge dx_n$ asigna a cada paralelepípedo orientado F de E el determinante de las coordenadas de sus vectores en la base $v_1, \ldots, v_n$. Si ésta es ortonormal y F no es degenerado entonces $(dx_1 \wedge \cdots \wedge dx_n)(F)$ es lo que llamábamos la medida orientada de F, es decir, la medida de $P(F)$ salvo un signo, que será positivo o negativo según la orientación de F.

Si la base $v_1, \ldots, v_n$ no es ortonormal (lo cual será lo más frecuente, pues las bases asociadas a cartas en espacios tangentes casi nunca lo son) entonces el volumen de F no vendrá dado por la n-forma básica, sino por un múltiplo suyo:

$$dm = a\,dx_1 \wedge \cdots \wedge dx_n,$$

donde a es el valor absoluto del determinante de la matriz de cambio de base entre una base ortonormal de E y $(v_1, \ldots, v_n)$. A esta n-forma se le llama *elemento de longitud, área, volumen, medida* de E (según la dimensión). Notemos que si hacemos actuar los dos miembros de la igualdad anterior sobre la base $(v_1, \ldots, v_n)$ vemos que a no es sino la medida del paralelepípedo que ésta determina.

En el caso en que E es un espacio tangente $T_p(S)$ y la base fijada es la asociada a una carta X, el teorema 9.2 prueba de hecho que

$$dm(p) = \Delta_X(p)\,dx_1(p) \wedge \cdots \wedge dx_n(p). \tag{9.2}$$

En efecto, basta observar que $\Delta_X(p)$ es la medida de la imagen por $dX(x)$ del paralelepípedo asociado a la base canónica, es decir, del paralelepípedo $(D_1X(x), \ldots, D_nX(x))$ que hemos tomado como base en $T_p(S)$.

Ahora nos encontramos con un inconveniente: nos gustaría definir el elemento de medida de una variedad S como la forma diferencial dm que a cada punto p le asigna la medida orientada de $T_p(S)$. Sin embargo esto es ambiguo, pues en $T_p(S)$ hay dos medidas orientadas —de signo opuesto— correspondientes a las dos orientaciones posibles del espacio. Más exactamente, si consideramos la expresión (9.2) para dos cartas distintas X e Y alrededor de un punto p, las formas $dm(p)$ correspondientes pueden ser iguales u opuestas según lo sean las orientaciones de las bases de $T_p(S)$ asociadas a las cartas. En otras palabras, depende de si $dX(x)$ y $dY(y)$ transforman la base canónica de $\mathbb{R}^n$ en bases con

la misma orientación, y es fácil ver que esto equivale a que $d(X \circ Y^{-1})(x)$ conserve la orientación, es decir, a que el determinante jacobiano de $X \circ Y^{-1}$ sea positivo en x.

Definición 9.14 Un *atlas* de una variedad S es un conjunto de cartas que cubran todos los puntos de S. Un *atlas orientado* es un atlas de S tal que si X e Y son dos de sus cartas y ambas cubren a un punto p, entonces el determinante jacobiano de $X \circ Y^{-1}$ es positivo. Una variedad S es *orientable* si admite un atlas orientado.

De este modo, si fijamos un atlas orientado en una variedad S y tomamos $p \in T_p(S)$, todas las bases de $T_p(S)$ inducidas por cartas del atlas tienen la misma orientación, a la que llamaremos *orientación positiva* de $T_p(S)$. En lo sucesivo, cuando hablemos de una variedad orientable se sobrentenderá que en ella hemos seleccionado un atlas orientado y por consiguiente una orientación positiva en cada espacio tangente.

Ejemplo Toda variedad cubrible por una sola carta es orientable, considerando el atlas formado únicamente por dicha carta. En particular todo abierto de $\mathbb{R}^n$ es una variedad orientable, tomando como carta la identidad. En lo sucesivo consideraremos siempre esta orientación en los abiertos de $\mathbb{R}^n$, de modo que la base canónica será una base orientada de cada espacio tangente. ■

Teorema 9.15 *Sea S una variedad orientable y X una carta de S con imagen conexa. Entonces, o bien las bases asociadas a X son todas positivas o bien son todas negativas. Según el caso diremos que la carta es positiva o negativa.*

Demostración: Sea U el dominio de X. Observamos que el conjunto de los puntos $x \in U$ tales que la orientación de la base de $T_{X(x)}(S)$ es positiva es un abierto. En efecto, si x es uno de estos puntos e Y es una carta del atlas orientado alrededor de $X(p)$, entonces el determinante jacobiano de $X \circ Y^{-1}$ es positivo en x, luego es positivo en un entorno de x, y en todos los puntos de dicho entorno la base asociada a X será positiva.

Similarmente se prueba que el conjunto de puntos $x \in U$ tales que la orientación de la base de $T_{X(x)}(S)$ es negativa es un abierto. Como U es conexo, uno de los dos conjuntos es vacío. ■

Casi todas las variedades que hemos considerado como ejemplos concretos son orientables. Las únicas excepciones son la banda de Möbius, que describimos brevemente en el capítulo V y el *plano elíptico, que contiene una banda de Möbius. Se puede probar que toda variedad de dimensión 1 es orientable. Es fácil ver que también lo es toda superficie de revolución, todo producto de variedades orientables (tomando como cartas positivas los productos de cartas positivas) así como la esfera, lo cual se deduce fácilmente a partir del teorema siguiente:

Teorema 9.16 *Una variedad $S \subset \mathbb{R}^m$ de dimensión $m-1$ es orientable si y sólo si existe una determinación continua $n : S \longrightarrow \mathbb{R}^m$ del vector normal a $T_p(S)$ en cada punto $p \in S$.*

DEMOSTRACIÓN: Si S es orientable y $p \in S$ definimos $n(p)$ como el único vector unitario tal que si X es una carta positiva alrededor de p la base de $\mathbb{R}^m$ formada por $\big(n(p), D_1X(x_0), \ldots, D_{m-1}X(x_0)\big)$ es positiva, donde x_0 es el vector de coordenadas de p. Es claro que $n(p)$ no depende de la elección de X. Veamos que n es diferenciable en un entorno de p, para lo cual probaremos que $X \circ n$ es diferenciable en un entorno de x_0. Consideremos un vector de indeterminadas $\overline{n} = (y_1, \ldots, y_m) \in \mathbb{R}^m$ y las siguientes m ecuaciones con variables $x_1, \ldots, x_{m-1}, y_1, \ldots, y_m$:

$$\overline{n}\, D_1X = 0, \ \ldots, \ \overline{n}\, D_{m-1}X = 0, \ \|\overline{n}\|^2 = 1.$$

Admitiendo que X es de clase C^2, el teorema de la función implícita nos da que en un entorno de x_0 existe una función diferenciable $\overline{n}$ tal que $\overline{n}(x) = n(p)$ y $\overline{n}(x)$ es unitario y perpendicular a $T_{X(x)}(S)$. En efecto, el determinante que ha de ser no nulo para ello está formado por los $m-1$ vectores $D_iX(x_0)$ y el vector $2n(p)$, que forman una base de $\mathbb{R}^m$.

Hemos de comprobar que $\overline{n}(x) = n(X(x))$ para todo x en el dominio de $\overline{n}$. Basta ver que el determinante cuyas filas son $\big(\overline{n}(x), D_1X(x), \ldots, D_{m-1}X(x)\big)$ es positivo en todo punto x, pero ciertamente es una función continua que no se anula y en x_0 es positivo, luego lo es en todos los puntos.

Recíprocamente, si n es una determinación continua del vector normal a S, definimos las cartas positivas como aquellas cartas X tales que la base $\big(n(X(x)), D_1X(x), \ldots, D_{m-1}X(x)\big)$ es positiva en todo punto x. Por el mismo argumento que en la implicación anterior es claro que la orientación de dicha base depende sólo de X. Si una carta X es negativa, entonces la carta $X(-x_1, x_2, \ldots, x_{m-1})$ es positiva y cubre los mismos puntos, luego todo punto de S se puede cubrir por una carta positiva. Es claro que si X e Y son dos cartas positivas alrededor de un punto p, entonces sus base asociadas en $T_p(S)$ tienen la misma orientación, luego las cartas positivas forman realmente un atlas orientado de S. ∎

La prueba del teorema anterior muestra que las determinaciones continuas del vector normal son de hecho diferenciables (si las cartas son de clase C^k, la aplicación n es de clase C^{k-1}).

Si una superficie encierra una región del espacio (una esfera, un cilindro, un toro, etc.) es costumbre tomar como orientación positiva la inducida por la determinación del vector normal que apunta *hacia fuera.*

Definición 9.17 El *elemento de medida* de una variedad orientable S es la forma diferencial dm que a cada punto $p \in S$ le asigna la medida orientada en $T_p(S)$, entendiendo que los paralelepípedos positivamente orientados tienen medida positiva. Si X es una carta positiva alrededor de un punto p, el elemento de medida en p viene dado por la expresión (9.2), que prueba que efectivamente es una forma de clase C^∞.

Esto nos da una caracterización de las variedades orientables en términos de formas diferenciales:

Teorema 9.18 *Una variedad diferencial S de dimensión n es orientable si y sólo si existe una n-forma $\omega \in \Lambda^n(S)$ tal que $\omega(p) \neq 0$ para todo $p \in S$.*

DEMOSTRACIÓN: Si S es orientable, entonces el elemento de medida dm es una n-forma en S que no se anula. Si existe una forma ω que no se anula, basta tomar como cartas positivas a las cartas X tales que para todo punto p cubierto por X se cumpla

$$\omega(p) = a(p)\,dx_1(p) \wedge \cdots \wedge dx_n(p), \quad \text{con } a > 0.$$

Notemos que si una carta es negativa, la carta que resulta de cambiar el signo a una función coordenada resulta positiva, por lo que todo punto puede cubrirse con una carta positiva, es decir, las cartas positivas forman un atlas. Por otra parte, si X e Y son dos cartas positivas alrededor de un punto p, tenemos que

$$\omega(p) = a(p)\,dx_1(p) \wedge \cdots \wedge dx_n(p) = b(p)\,dy_1(p) \wedge \cdots \wedge dy_n(p),$$

donde a y b son funciones positivas. Por el teorema 9.12 resulta que ba^{-1} es el determinante de la matriz de cambio de base entre $DX_1(p), \ldots, DX_n(p)$ y $DY_1(p), \ldots, DY_n(p)$, luego éste es positivo y ambas bases tienen la misma orientación, luego las cartas positivas forman un atlas orientado. ■

Definición 9.19 Sea S una variedad orientable de dimensión n y sea ω una n-forma diferencial en S. Entonces $\omega = f\,dm$, para cierta $f \in \Lambda^0(S)$. Diremos que ω es *integrable* en un conjunto de Borel $B \subset S$ si lo es f respecto a la medida de Lebesgue, y en tal caso definimos

$$\int_B \omega = \int_B f\,dm,$$

donde el segundo miembro se entiende como la integral de f respecto a la medida de Lebesgue de S.

Observar que si consideramos la integral de ω como función de B, tenemos que cada n-forma integrable induce una medida signada en S.

Definimos el *soporte* de una forma en S como la clausura del conjunto de puntos de S donde ω no es nula. Es claro que las n-formas continuas con soporte compacto son integrables.

Conviene comprender el significado geométrico de la medida asociada a una n-forma ω en una variedad S de dimensión n: es claro que ω determina una medida en cada espacio $T_p(S)$, la única medida que sobre los paralepípedos actúa como $\omega(p)$. Entonces, la medida que ω induce en S es la única medida que en un entorno de cada punto p se confunde con la medida correspondiente de $T_p(S)$, donde "se confunde" hay que entenderlo exactamente en el mismo sentido que al principio del capítulo.

Continuemos con las propiedades de las integrales de formas: Es claro que

$$\int_B (\alpha\omega + \beta\omega') = \alpha \int_B \omega + \beta \int_B \omega'.$$

Observar que si B está cubierto por una carta positiva X, aplicando la fórmula (9.2) y el teorema 9.3 tenemos

$$\int_B f\,dx_1\wedge\cdots\wedge dx_n = \int_B \frac{f}{\Delta_X}\,dm = \int_{X^{-1}[B]} (X\circ f)\,dx_1\cdots dx_n, \tag{9.3}$$

donde la última integral es una integral usual en $\mathbb{R}^n$ respecto a la medida de Lebesgue. En la práctica, si expresamos f en función de las coordenadas respecto a X, los dos extremos de las igualdades anteriores se escriben igual (salvo por la supresión de los símbolos del producto exterior), con lo que la teoría que subyace en dicha igualdad se vuelve "trasparente". El ejemplo siguiente ilustra lo que queremos decir:

Ejemplo Sea S la esfera de centro 0 y radio 1 (con la orientación usual, es decir, de modo que las bases positivas induzcan el vector normal que apunta hacia fuera). Sea B la semiesfera $z > 0$ y calculemos

$$\int_B xy\,dx\wedge dy.$$

Para ello observamos que la carta de coordenadas (x, y), es decir, la carta $X(x,y) = \bigl(x, y, \sqrt{1-x^2-y^2}\,\bigr)$ es positiva, pues

$$X_x\wedge X_y = \Bigl(\frac{x}{z}, \frac{y}{z}, 1\Bigr) = \frac{1}{z}(x,y,z).$$

Por consiguiente

$$\int_B xy\,dx\wedge dy = \int_C xy\,dxdy,$$

donde $C = \{(x,y)\in\mathbb{R}^2 \mid x^2+y^2<1\}$. Así pues (una vez hemos comprobado que la carta está bien orientada) el cálculo de la integral se reduce a considerar que xy no es la función que a cada punto de S le asigna el producto de sus coordenadas, sino simplemente la función xy en $\mathbb{R}^2$, y que éstas varían en el conjunto en el que varían las coordenadas de los puntos de la semiesfera, es decir, en el disco unitario. ■

Ejemplo Sea de nuevo S la esfera unidad. Veamos que

$$\int_S dx\wedge dy = 0.$$

Llamemos B^+ y B^- a las semiesferas $z > 0$ y $z < 1$. Teniendo en cuenta que el ecuador tiene medida nula, podemos escribir

$$\int_S dx\wedge dy = \int_{B^+} dx\wedge dy + \int_{B^-} dx\wedge dy.$$

Razonando como en el ejemplo anterior concluimos que la primera integral vale π, el área del disco unitario de $\mathbb{R}^2$. Para calcular la integral en la otra

semiesfera no podemos usar la carta $X(x,y) = \big(x, y, -\sqrt{1-x^2-y^2}\,\big)$ porque es negativa. En su lugar usamos $X(y,x) = \big(x, y, -\sqrt{1-x^2-y^2}\,\big)$, con lo que

$$\int_{B^-} dx \wedge dy = -\int_{B^-} dy \wedge dx = -\int_C dydx = -\pi,$$

luego la integral total es nula. ■

Para trabajar teóricamente con transformaciones de integrales del estilo de las que hemos empleado en los ejemplos anteriores conviene introducir un nuevo concepto.

Teorema 9.20 *Sea $f : S \longrightarrow T$ una aplicación diferenciable entre variedades. Entonces f induce un homomorfismo de álgebras $f^\sharp : \Lambda(T) \longrightarrow \Lambda(S)$ que a cada $\omega \in \Lambda^k(T)$ le asigna la k-forma dada por*

$$f^\sharp(\omega)(p)(v_1, \ldots, v_k) = \omega\big(f(p)\big)\big(df(p)(v_1), \ldots, df(p)(v_k)\big).$$

Demostración: Se comprueba inmediatamente que $f^\sharp(\omega)(p) \in A^k\big(T_p(S)\big)$, con lo que $f^\sharp$ es una aplicación de $\Lambda(T)$ en el álgebra de todas las formas de S (no necesariamente diferenciables). Así mismo es claro que $f^\sharp$ es lineal y, a partir de la definición del producto exterior, se comprueba también sin dificultad que $f^\sharp(\omega \wedge \omega') = f^\sharp(\omega) \wedge f^\sharp(\omega')$.

Para probar que $f^\sharp(\omega)$ es diferenciable en un punto p tomamos una carta $X : U \longrightarrow V$ alrededor de p y una carta $Y : U' \longrightarrow V'$ alrededor de $f(p)$. Podemos suponer que $f[V] \subset V'$. Basta ver que $f^\sharp(\omega)|_V$ es diferenciable en p, pero como la definición de $f^\sharp$ depende sólo de los valores que toma f alrededor de p, dicha restricción coincide con $(f|_V)^\sharp(\omega|_{V'})$. En resumen, que podemos suponer que S y T son simplemente V y V'.

En tal caso, las diferenciales dy_i junto con las 0-formas generan toda el álgebra de Grassmann, luego basta probar que $f^\sharp(dy_i)$ y $f^\sharp(g)$ con $g \in \Lambda^0(T)$ son diferenciables. Esto es evidente: por la propia definición $f^\sharp(dy_i) = d(f \circ y_i)$ y $f^\sharp(g) = f \circ g$. ■

Una simple comprobación nos da que $(f \circ g)^\sharp = g^\sharp \circ f^\sharp$. Si I es la identidad en una variedad S, entonces $I^\sharp$ es la identidad en $\Lambda(S)$, luego si f es un difeomorfismo también lo es $f^\sharp$ y $(f^\sharp)^{-1} = (f^{-1})^\sharp$.

La aplicación $f^\sharp$ recibe el nombre de *retracción* asociada a f.

Cuando se emplea la notación adecuada, las retracciones resultan "invisibles" en la práctica. Por ejemplo, la retracción de la inclusión $i : S \longrightarrow T$ entre dos variedades es $i^\sharp(\omega)(p) = \omega(p)|_{T_p(S)^k}$, donde $\omega \in \Lambda^k(T)$. En muchas ocasiones hemos usado la notación dx_i para deferirnos tanto a la diferencial en $\mathbb{R}^n$ de la proyección x_i en la i-ésima componente como para referirnos a la diferencial de la restricción de x_i a una variedad S. Ahora vemos que dicha restricción es $i^\sharp(x_i)$ y que la diferencial de la restricción es $i^\sharp(dx_i)$.

Otro ejemplo nos lo proporciona la fórmula (9.3), que hemos usado para calcular integrales en variedades. En términos de retracciones se escribe como

$$\int_B \omega = \int_{X^{-1}[B]} X^\sharp(\omega),$$

donde X es un carta de una variedad y B es un conjunto de Borel en su rango.

En efecto, notemos que las x_i del segundo miembro de (9.3) son en realidad $X \circ x_i$ (si entendemos que las x_i son, como en el primer miembro, las coordenadas de X en la variedad), luego $dx_i(u)(v)$ es en realidad $dx_i(X(u))(dX(v))$, es decir, las diferenciales dx_i que aparecen en el segundo miembro son en realidad $X^\sharp(dx_i)$, si entendemos dx_i como en el primer miembro.

Con rigor deberíamos escribir $X^\sharp(i^\sharp(\omega))$, donde i es la inclusión del rango de X en la variedad. Usando particiones de la unidad podemos probar un resultado más general, pero primero necesitamos justificar que las particiones se pueden tomar de clase C^∞. Ello se debe al teorema siguiente, que es la versión del Lema de Urysohn para funciones de clase C^∞. Recordemos que la notación $K \prec f \prec V$ en un espacio S significa que $f : S \longrightarrow [0,1]$ es una aplicación continua que vale 1 en K y se anula fuera de V. En lo sucesivo S será una variedad y $K \prec f \prec V$ supondrá también que f es de clase C^∞.

Teorema 9.21 *Sea S una variedad diferenciable, sea K un subconjunto compacto de S y sea V un abierto de modo que $K \subset V$. Entonces existe $f \in \Lambda^0(S)$ tal que $K \prec f \prec V$.*

DEMOSTRACIÓN: Probamos primero que dados números reales $0 \leq a < b$ existe $g : \mathbb{R}^n \longrightarrow [0,1]$ de clase C^∞ tal que

$$g(x) = \begin{cases} 0 & \text{si } \|x\| \leq a, \\ 1 & \text{si } \|x\| \geq b. \end{cases}$$

La construcción es sencilla a partir del teorema 3.32. En efecto, según dicho teorema existe una función $f : \mathbb{R} \longrightarrow \mathbb{R}$ de clase C^∞ que es nula fuera del intervalo $\left]a^2, b^2\right[$ y positiva en su interior. Sea $M = \int_{a^2}^{b^2} f(t)\,dt > 0$. Definimos

$$\phi(u) = \frac{1}{M}\int_{a^2}^{u} f(t)\,dt.$$

Así, $M\phi(u)$ es una primitiva de f, luego es de clase C^∞ en $\mathbb{R}$ y ϕ también. Como f es mayor o igual que 0 es claro que ϕ es creciente. También es obvio que

$$\phi(u) = \begin{cases} 0 & \text{si } u \leq a^2, \\ 1 & \text{si } u \geq b^2. \end{cases}$$

Ahora basta tomar $g(x) = 1 - \phi(\|x\|^2)$.

Sea ahora $p \in K$. Tomamos una carta $X_p : U \longrightarrow W \subset S$ que cubra a p, donde U es un abierto en $\mathbb{R}^n$. Podemos suponer que $0 \in U$, $X_p(0) = p$ y $W \subset V$. Sea $b_p > 0$ tal que $\overline{B_{b_p}(0)} \subset U$. Sea $a_p = b_p/2$. Componiendo X_p^{-1} con la función que nos da la construcción anterior obtenemos una función $g_p : W \longrightarrow [0,1]$ de clase C^∞ en W tal que g_p vale 1 en un entorno de p en S (concretamente en $X_p[B_{a_p}(0)]$) y se anula fuera de $X_p[\overline{B_{b_p}(0)}]$. Es claro que podemos extenderla a una función $g_p : S \longrightarrow [0,1]$ de clase C^∞ sin más que hacer $g_p(q) = 0$ para todo $q \in S \setminus W$. En particular g_p se anula fuera de V.

Hemos dicho que g_p vale 1 en un entorno de p. Estos entornos cubren K, luego por compacidad podemos extraer un subcubrimiento finito, con lo que tenemos k abiertos $W_1, \ldots, W_k$ que cubren K y k funciones $g_i : S \longrightarrow [0, 1]$ de clase C^∞ de modo que se anulan fuera de V y $g_i|_{W_i} = 1$.

Ahora basta tomar $f(q) = 1 - (1 - g_1(q)) \cdots (1 - g_k(q))$. ■

Notemos que el teorema 7.27 es válido para variedades diferenciables, pues éstas son localmente compactas y además, si en lugar de usar en la prueba el Lema de Urysohn usamos el teorema 9.21, resulta que las particiones de la unidad las podemos tomar de clase C^∞, tal y como pretendíamos. Ahora ya podemos probar:

Teorema 9.22 *Sea $f : S \longrightarrow T$ un difeomorfismo entre variedades orientables de dimensión n y ω una n-forma en T con soporte compacto. Entonces*

$$\int_T \omega = \int_S f^\sharp(\omega).$$

DEMOSTRACIÓN: Cubrimos el soporte de ω con un número finito de rangos de cartas. Tomamos una partición de la unidad $h_1, \ldots, h_r$ subordinada a tales abiertos, es decir, $h_1 + \cdots h_r = 1$ sobre los puntos del soporte de ω y cada h_i tiene su soporte contenido en el rango de una carta. Entonces $\omega = h_1\omega + \cdots + h_r\omega$ y basta probar el teorema para cada forma $h_i\omega$. Equivalentemente, podemos suponer que el soporte de ω está contenido en el rango V de una carta X de T. Entonces $Y = X \circ f^{-1}$ es una carta de S y es fácil ver que el soporte de $f^\sharp(\omega)$ está contenido en su rango. Según los comentarios previos al teorema, las integrales de la igualdad que queremos probar coinciden respectivamente con las de las formas $Y^\sharp(f^\sharp(\omega))$ y $X^\sharp(\omega)$, definidas sobre el dominio (común) de las dos cartas. Es claro que ambas formas son la misma. ■

Veamos otro ejemplo de "invisibilidad" de las retracciones: Si S_1 y S_2 son dos variedades diferenciables, la retracción de la proyección $\pi_i : S_1 \times S_2 \longrightarrow S_i$ es un homomorfismo $\pi_i^\sharp : \Lambda(S_i) \longrightarrow \Lambda(S_1 \times S_2)$.

Supongamos que S_1 y S_2 son los rangos de las cartas X_1 y X_2, de coordenadas $x_1, \ldots, x_{n_1}$ e $y_1, \ldots, y_{n_2}$. Entonces las funciones coordenadas de $X_1 \times X_2$ son las $\pi_1 \circ x_i$ y $\pi_2 \circ y_i$, que en la práctica podemos llamar también x_i e y_i, pero que con rigor son $\pi_1^\sharp(x_i)$ y $\pi_2^\sharp(y_i)$.

Notemos que $d\pi_i(p_1, p_2) : T_{p_1}(S_1) \times T_{p_2}(S_2) \longrightarrow T_{p_i}(S_i)$ es simplemente la proyección. Teniendo esto en cuenta es fácil ver que dx_i, considerada como forma en $S_1 \times S_2$, no es sino la retracción $\pi_1^\sharp(dx_i)$, donde ahora dx_i es la forma de S_1. Así pues, la retracción de una forma arbitraria de S_i expresada en términos de las coordenadas de X_i es la forma que tiene la misma expresión pero interpretando las coordenadas y las diferenciales en el producto.

Más en general, es fácil ver que si una k-forma ω (digamos de S^1) no se anula en un punto p, entonces $\pi_1^\sharp(\omega)$ no se anula en los puntos de la forma (p, q) y análogamente para $i = 2$, con lo que las retracciones $\pi_i^\sharp$ son monomorfimos de álgebras y podemos identificar las formas de S_1 y S_2 con formas de $S_1 \times S_2$.

Por ejemplo, con estas identificaciones tenemos que $dm = dm_1 \wedge dm_2$, donde dm, dm_1 y dm_2 son los elementos de medida de $S_1 \times S_2$, S_1 y S_2 respectivamente. En efecto, una carta $X_1 \times X_2$ alrededor de un punto (p, q) de coordenadas (x, y) induce la base de $T_{(p,q)}(S_1 \times S_2)$ formada por los vectores $(D_iX_1(x), 0)$ y $(0, D_iX_2(y))$. Además

$$u_i = d\pi_1(p, q)(D_iX_1(x), 0) = D_iX_1(x), \quad d\pi_1(p, q)(0, D_iX_2(y)) = 0,$$

$$v_i = d\pi_2(p, q)(0, D_iX_2(y)) = D_iX_2(y), \quad d\pi_2(p, q)(D_iX_1(x), 0) = 0,$$

luego, al calcular $\pi_1^\sharp(dm_1)(p) \wedge \pi_2^\sharp(dm_2)(q)$ sobre esta base mediante la definición de producto exterior, se anulan todos los sumandos correspondientes a permutaciones que hacen actuar a $\pi_1^\sharp(dm_1)(p)$ sobre una vector de X_2 y viceversa. Por consiguiente queda

$$\sum_{(\sigma,\tau)\in\Sigma_{n_1}\times\Sigma_{n_2}} \frac{\operatorname{sig}(\sigma,\tau)}{n_1!n_2!}\, dm_1(u_{\sigma(1)}, \ldots, u_{\sigma(n_1)})\, dm_2(v_{\tau(1)}, \ldots, v_{\tau(n_2)}),$$

que claramente es igual a

$$dm_1(p)(u_1, \ldots, u_{n_1})dm_2(q)(v_1, \ldots, v_{n_2}) = \Delta_{X_1}(p)\Delta_{X_2}(q) = \Delta_{X_1\times X_2}(p, q).$$

Por otra parte es inmediato que éste es el valor que toma $dm(p, q)$ sobre la misma base, luego ambas formas coinciden. ■

Ejemplo La aplicación $f : \,]0, +\infty[\, \times S^{n-1} \longrightarrow \mathbb{R}^n \setminus \{0\}$ dada por $f(r, x) = rx$ es un difeomorfismo. Tomemos una carta X de S^{n-1} (de coordenadas $x_1, \ldots, x_{n-1}$) y la identidad como carta de $]0, +\infty[$ (con coordenada r). Entonces $Y = (I \times X) \circ f$ es una carta de $\mathbb{R}^n \setminus \{0\}$. Más detalladamente: $Y(r, x_1, \ldots, x_{n-1}) = rX(x_1, \ldots, x_{n-1})$.

De este modo, cada punto del rango de $I \times X$ tiene las mismas coordenadas que su imagen por f y la retracción $f^\sharp$ de una forma de $\mathbb{R}^n \setminus \{0\}$ expresada en términos de $r, x_1, \ldots, x_{n-1}$, y sus diferenciales es la forma con la misma expresión pero interpretada como forma de $]0, +\infty[\, \times S^{n-1}$.

En estas coordenadas, el elemento de medida en $\mathbb{R}^n \setminus \{0\}$ es

$$dm = r^{n-1} \begin{vmatrix} X_1 & D_1X_1 & \cdots & D_{n-1}X_1 \\ \vdots & \vdots & & \vdots \\ X_n & D_1X_n & \cdots & D_{n-1}X_n \end{vmatrix} dr \wedge dx_1 \wedge \cdots \wedge dx_{n-1}.$$

La retracción $f^\sharp(dm)$ viene dada por esta misma expresión. El determinante es la medida del paralelepípedo formado por los vectores $X, D_1X, \ldots, D_{n-1}X$. Como X es unitario y perpendicular a los otros vectores, es claro que dicha medida es también la medida $n-1$-dimensional del paralelepípedo determinado por los vectores $D_1X, \ldots, D_{n-1}X$. Así pues,

$$f^\sharp(dm) = r^{n-1}\, dr \wedge d\sigma,$$

donde $d\sigma$ es el elemento de medida de S^{n-1}. Hemos probado esta relación para los puntos de $]0,+\infty[\times S^{n-1}$ cubiertos por la carta que hemos tomado, pero como ésta era arbitraria, la igualdad vale para todo punto. Por otro lado $dr\wedge d\sigma$ es el elemento de medida de $]0,+\infty[\times S^{n-1}$.

Con esto hemos probado en general que si h es una función integrable en $\mathbb{R}^n$, entonces

$$\int_{\mathbb{R}^n} h(r,x_1,\ldots,x_{n-1})\,dm = \int_{]0,+\infty[\times S^{n-1}} h(r,x_1,\ldots,x_{n-1})r^{n-1}\,dm$$

$$= \int_0^{+\infty}\left(\int_{S^{n-1}} h(r,x_1,\ldots,x_{n-1})r^{n-1}\,d\sigma\right)dr.$$

De aquí se sigue que si identificamos $\mathbb{R}^n$ con $]0,+\infty[\times S^{n-1}$ entonces la medida de Lebesgue se identifica con el producto de la medida dada por $r^{n-1}\,dr$ y la medida de Lebesgue en S^{n-1}. En particular, podemos aplicar el teorema de Fubini, que nos asegura que si h es positiva y existe la integral doble del miembro derecho, entonces h es integrable en $\mathbb{R}^n$. En otras palabras, las funciones pueden integrarse "por capas". ■

Veamos un par de aplicaciones:

Teorema 9.23 *La función* $1/\|x\|^\alpha$ *es integrable en un entorno de* 0 *en* $\mathbb{R}^n$ *si y sólo si* $\alpha < n$.

Demostración: Con la notación del ejemplo anterior:

$$\int_{D(\epsilon,R)} \frac{1}{\|x\|^\alpha}\,dm = \int_{]\epsilon,R[\times S^{n-1}} r^{-\alpha}r^{n-1}\,dm = m(S^{n-1})\int_\epsilon^R r^{n-1-\alpha}\,dr.$$

Si $\alpha < n$, una primitiva del integrando de la derecha es $r^{n-\alpha}/(n-\alpha)$ y podemos concluir que

$$\int_B \frac{1}{\|x\|^\alpha}\,dm = m(S^{n-1})\frac{R^{n-\alpha}}{n-\alpha}.$$

Es fácil ver que si $\alpha \geq n$ el miembro derecho tiende a ∞ cuando ϵ tiende a 0. ■

Ejemplo Sea σ_n la medida de Lebesgue de S^n. Vamos a calcularla. Para ello consideremos la función $g(x) = e^{-\|x\|^2}$, definida en $\mathbb{R}^{n+1}$. Calculamos de dos formas su integral. Por una parte (ver el final del capítulo anterior)

$$\int_{\mathbb{R}^{n+1}} g(x)\,dm = \left(\int_{\mathbb{R}} e^{-t^2}\,dt\right)^{n+1} = \pi^{(n+1)/2}.$$

Por otra parte la integral se puede calcular como

$$\int_{S^n}\int_0^{+\infty} r^n e^{-r^2}\,dr d\sigma = \frac{\sigma_n}{2}\int_0^{+\infty} t^{n/2}e^{-t}t^{-1/2}\,dt = \frac{\sigma_n}{2}\Pi(\frac{n-1}{2}).$$

Por consiguiente

$$\sigma_n = \frac{2\pi^{(n+1)/2}}{\Pi(\frac{n-1}{2})} = \frac{(n+1)\pi^{(n+1)/2}}{\Pi(\frac{n+1}{2})}.$$

En el capítulo anterior obtuvimos una expresión para la medida de Lebesgue de la bola unidad en términos de la función factorial. También podemos deducirla de la expresión anterior, pues a través de la identificación $\mathbb{R}^n =]0, +\infty[\times S^{n-1}$ la bola unidad se identifica con $]0, 1[\times S^{n-1}$ y su medida de Lebesgue es el producto de la medida $r^{n-1}\,dr$ por la medida de Lebesgue en la esfera. La medida del intervalo es

$$\int_0^1 r^{n-1}\,dr = \frac{1}{n},$$

luego la medida de la bola es

$$v_n = \frac{\sigma_{n-1}}{n} = \frac{\pi^{n/2}}{\Pi(n/2)},$$

como ya sabíamos. ■

9.4 Algunos conceptos del cálculo vectorial

En esta sección introduciremos algunos conceptos que ilustren las aplicaciones de las formas diferenciales. Comenzamos con la integral curvilínea y sus aplicaciones. Sea $\gamma : [a, b] \longrightarrow \mathbb{R}^n$ un arco regular. Si γ no se corta a sí mismo (sin excluir que sus extremos coincidan) podemos considerar a su imagen como una 1-variedad con γ como única carta y dotada de la orientación ésta le induce. Sea $F : \gamma[a, b] \longrightarrow \mathbb{R}^n$ un campo continuo de vectores. A este campo le asociamos la 1-forma en γ dada por

$$F\,d\vec{r} = F_1\,dx_1 + \cdots + F_n\,dx_n,$$

que formalmente puede interpretarse como el producto escalar de F por el vector $d\vec{r} = (dx_1, \ldots, dx_n)$. De hecho, la forma asigna a cada vector $v \in T_p(\gamma)$ el producto escalar $F(p)v$. Esto implica que $F\,d\vec{r}$ sólo depende de la proyección de F sobre la recta tangente a γ en cada punto. En particular, si F es el vector tangente unitario T entonces $(T\,d\vec{r})(v) = \pm\|v\|$, la longitud orientada de v, es decir, $T\,d\vec{r} = ds$. En general podemos descomponer $F = aT + N$, donde T y N son ortogonales. Multiplicando por T vemos que $a = FT$. Así pues $F\,d\vec{r} = (FT)T\,d\vec{r} = FT\,ds$.

Se define la *circulación* o *integral curvilínea* de F a través de γ como la integral

$$\begin{aligned}\int_\gamma F\,d\vec{r} &= \int_\gamma FT\,ds = \int_a^b \big(F_1\big(\gamma(t)\big)\gamma_1'(t) + \cdots + F_n\big(\gamma(t)\big)\gamma_n'(t)\big)dt \\ &= \int_a^b F\big(\gamma(t)\big)\gamma'(t)\,dt.\end{aligned}$$

Observemos que $F\,d\vec{r}$ es en principio una forma continua, no necesariamente diferenciable, pero nuestra definición de integral vale igualmente en este caso. Notemos también que la última integral tiene sentido aunque γ se corte a sí misma, por lo que la noción de integral curvilínea es ligeramente más general. En la práctica conviene considerar incluso el caso en que γ es derivable salvo en un número finito de puntos, lo cual tampoco afecta a la integral.

Definición 9.24 Un *arco singular* es una aplicación continua $\phi : [a, b] \longrightarrow \mathbb{R}^n$ tal que existe una partición $a = t_0 < t_1 < \cdots < t_n = b$ de modo que $\phi|_{[t_i,t_{i+1}]}$ es de clase C^1 (en el sentido de que se extiende a una función de clase C^1 en un abierto que contiene a $[t_i, t_{i+1}]$).

Observar que no exigimos que la derivada no se anule. Si $F : \phi[a, b] \longrightarrow \mathbb{R}^n$ tenemos definida la integral curvilínea de F sobre ϕ mediante

$$\int_\phi F\,d\vec{r} = \sum_{i=1}^n \int_{t_{i-1}}^{t_i} F\big(\phi(t)\big)\phi'(t)\,dt.$$

Si $\phi : [a, b] \longrightarrow \mathbb{R}^n$ y $\psi : [c, d] \longrightarrow \mathbb{R}^n$ son dos arcos singulares tales que $\phi(b) = \psi(c)$, definimos su *unión* como el arco $\phi \cup \psi : [a, b + d - c] \longrightarrow \mathbb{R}^n$ dado por

$$(\phi \cup \psi)(t) = \begin{cases} \phi(t) & \text{si } a \leq t \leq b \\ \psi(t - b + c) & \text{si } b \leq t \leq b + d - c \end{cases}$$

Claramente $\phi\cup\psi$ es el arco que resulta de recorrer ϕ y seguidamente recorrer ψ. Observemos que aunque ϕ y ψ fueran derivables en todo su dominio, su unión no tiene por qué serlo en el punto de enlace. Ésta es una de las razones por las que conviene trabajar con arcos derivables a trozos.

Dado un arco singular $\phi : [a, b] \longrightarrow \mathbb{R}^n$, definimos su *inverso* como el arco $-\phi : [-b, -a] \longrightarrow \mathbb{R}^n$ dado por $(-\phi)(t) = \phi(-t)$. Se trata del arco que recorre la misma trayectoria pero en sentido inverso.

Las propiedades siguientes son inmediatas:

$$\int_{\phi\cup\psi} F\,d\vec{r} = \int_\phi F\,d\vec{r} + \int_\psi F\,d\vec{r}, \qquad \int_{-\phi} F\,d\vec{r} = -\int_\phi F\,d\vec{r}.$$

En este contexto, el segmento que une dos puntos x, $y \in \mathbb{R}^n$ es el arco dado por $[x, y](t) = (1 - t)x + ty$, para $0 \leq t \leq 1$. Una poligonal es una unión finita de segmentos. Un arco $\phi : [a, b] \longrightarrow \mathbb{R}^n$ es *cerrado* si $\phi(a) = \phi(b)$.

La interpretación más importante de la circulación de un campo a lo largo de una trayectoria proviene de la física:

Ejemplo Supongamos que $\gamma(t)$ es la posición en cada instante t de un móvil de masa m, de modo que cuando se encuentra en la posición p la fuerza total que actúa sobre él es $F(p)$. La circulación W de F a través de γ recibe el nombre de *trabajo* realizado por F sobre el móvil. Para entender el significado físico del trabajo observemos en primer lugar que éste depende únicamente de de la

componente tangencial de F, que será de la forma $F_t = maT$, donde T es el vector tangente unitario de γ y a es la derivada del módulo v de la velocidad del móvil. Así

$$dW = F\,d\vec{r} = maT\,d\vec{r} = ma\,ds = mav\,dt = mv\,dv.$$

Nos gustaría integrar ambos miembros, pero para ello necesitaríamos considerar a v como variable independiente, lo que equivale a tomarla como parámetro de γ y esto no siempre será posible. Pese a ello, el resultado que se obtiene de integrar formalmente la igualdad anterior es correcto. Para probarlo definimos

$$E = \frac{1}{2}mv^2.$$

Entonces

$$\frac{dE}{dt} = mva = \frac{dW(t)}{dt},$$

donde $W(t)$ representa a la circulación de F en el intervalo $[a, t]$. De aquí se sigue que

$$W = W(b) = \Delta E = E(b) - E(a).$$

La magnitud E recibe el nombre de *energía cinética* del móvil, y lo que hemos probado es que el trabajo que ejerce una fuerza sobre un móvil es igual al incremento de la energía cinética que éste experimenta bajo su influencia. Notar que podemos definir el trabajo realizado por cualquier fuerza sobre un móvil dado, no necesariamente la fuerza total que actúa sobre él. Entonces el trabajo realizado por una suma de fuerzas es la suma de los trabajos realizados por cada una de ellas. El trabajo y la energía se miden en Julios. Un *Julio* es el trabajo que realiza una fuerza de un Newton cuando actúa tangencialmente sobre un móvil que recorre una trayectoria de un metro. ■

Existe una versión de la regla de Barrow para integrales curvilíneas.

Teorema 9.25 *Sea $f : U \longrightarrow \mathbb{R}$ una aplicación de clase C^2 en un abierto U de $\mathbb{R}^n$ y sea $\gamma : [a, b] \longrightarrow U$ es un arco de extremos $p = \gamma(a)$ y $q = \gamma(b)$. Entonces*

$$\int_\gamma df = f(q) - f(p).$$

Demostración: Observemos que

$$df = \frac{\partial f}{\partial x_1}\,dx_1 + \cdots + \frac{\partial f}{\partial x_n}\,dx_n = \nabla f\,d\vec{r}.$$

Entonces

$$\begin{aligned}\int_\gamma df &= \int_\gamma \nabla f\,d\vec{r} = \int_a^b \nabla f(\gamma(t))\gamma'(t)\,dt \\ &= \int_a^b \frac{d(\gamma\circ f)}{dt}\,dt = f\big(\gamma(b)\big) - f\big(\gamma(a)\big) = f(q) - f(p).\end{aligned}$$

(Si hay puntos donde γ no es derivable se razona separadamente en cada intervalo donde sí lo es y se llega a la misma conclusión). ■

Vemos así que cuando integramos un campo de la forma ∇f sobre un arco, la integral sólo depende de los extremos del mismo. Este hecho tiene gran importancia:

Definición 9.26 Diremos que un campo $F : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ definido en un abierto U es *conservativo* si la circulación de F a lo largo de cualquier arco singular contenido en U depende únicamente de sus extremos.

Hay que entender que en la dependencia de los extremos se incluye el orden de los mismos, pues γ y $-\gamma$ tienen los mismos extremos, pero las integrales respectivas son opuestas. Notar que si ϕ y ψ son arcos con los mismos extremos, la condición

$$\int_\phi F\,d\vec{r} = \int_\psi F\,d\vec{r}$$

es equivalente a

$$\int_{\phi\cup-\psi} F\,d\vec{r} = 0.$$

Es claro entonces que un campo F es conservativo si y sólo si las integrales de F a lo largo de los arcos cerrados son todas nulas.

El teorema anterior prueba que los campos de gradientes, es decir, los campos de la forma $F = \nabla f$, donde f es una función de clase C^2, son conservativos. Ahora probamos que éstos son los únicos campos conservativos:

Teorema 9.27 *Un campo $F : U \longrightarrow \mathbb{R}^n$ de clase C^1 en un abierto $U \subset \mathbb{R}^n$ es conservativo si y sólo si existe una función $V : U \longrightarrow \mathbb{R}$ tal que $F = \nabla V$. Si U es conexo, la función V está determinada salvo una constante.*

Demostración: Ya sabemos que los campos de gradientes son conservativos. Supongamos que F es un campo conservativo. No perdemos generalidad si suponemos que U es conexo. Entonces es conexo por poligonales. Fijamos un punto $x_0 \in U$ y para cada $x \in U$ existe una poligonal $\phi_x : [a, b] \longrightarrow U$ tal que $\phi_x(a) = x_0$ y $\phi_x(b) = x$. Definimos

$$V(x) = \int_{\phi_x} F\,d\vec{r}.$$

Como F es conservativo, $V(x)$ no depende de la elección de la poligonal. Veamos que $\nabla V = F$, lo que en particular probará que V es una función de clase C^2.

Tomemos $x \in V$ y sea e_i el i-ésimo vector de la base canónica de $\mathbb{R}^n$. Sea ϕ_x una poligonal que una x_0 con x y consideremos la poligonal $\phi_x \cup [x, x+he_i]$, que une x_0 con $x + he_i$, donde $h \neq 0$ es suficientemente pequeño para que $x + he_i$ esté en U. Entonces

$$\begin{aligned}\frac{V(x+he_i)-V(x)}{h} &= \frac{1}{h}\int_{[x,x+he_i]} F\,d\vec{r} = \frac{1}{h}\int_0^1 F(x+the_i)he_i\,dt \\ &= \int_0^1 F_i(x+the_i)\,dt\end{aligned}$$

La función F_i es uniformemente continua en el segmento $[x - h_0e_i, x + h_0e_i]$, para un h_0 fijo. Por lo tanto, dado $\epsilon > 0$, existe un $\delta > 0$ tal que si $|h| \leq \delta$ y $0 \leq t \leq 1$ entonces $|F_i(x + fhe_i) - F_i(x)| < \epsilon/2$. Por consiguiente

$$\left|\frac{V(x + he_i) - V(x)}{h} - F_i(x)\right| = \left|\int_0^1 \big(F_i(x + the_i) - F_i(x)\big)\, dt\right|$$

$$\leq \int_0^1 |F_i(x + the_i) - F_i(x)|\, dt \leq \frac{\epsilon}{2}\int_0^1 dt < \epsilon.$$

Esto prueba que existe

$$\frac{\partial V}{\partial x_i}(x) = F_i(x).$$

La unicidad de V es clara: si V_1 y V_2 cumplen el teorema entonces $V_1 - V_2$ es una función de clase C^1 con gradiente nulo, luego su diferencial es nula y (si U es conexo) $V_1 - V_2$ es constante. ■

Si un campo es de la forma $F = \nabla V$, se dice que la función V es una *función potencial* para F. Según hemos calculado, la circulación de F a lo largo de un arco singular es la diferencia de potencial entre sus extremos. De nuevo la física nos proporciona ejemplos de esta situación:

Ejemplo El campo gravitatorio que produce una masa puntual es conservativo. Recordemos que si el cuerpo tiene masa M y elegimos el sistema de referencia de modo que sus coordenadas sean nulas, la fuerza que éste ejerce sobre un cuerpo de masa m situado en la posición x es

$$F = -\frac{GMm}{\|x\|^3}\, x.$$

Puesto que la fuerza depende sólo de $\rho = \|x\|$ lo mismo ha de suceder con el potencial. Si planteamos el problema en una variable (y $x > 0$) la solución es simple: buscamos una función cuya derivada sea $-GMm/x^2$, luego nos sirve GMm/x. En tres dimensiones la solución es

$$\frac{GMm}{\|x\|}.$$

Si un cuerpo de masa m se encuentra en la posición x, definimos su *energía potencial* respecto a la masa M como

$$E_p = -\frac{GMm}{\|x\|}.$$

Añadimos el signo negativo de modo que si el cuerpo se desplaza desde un punto x hasta un punto y por cualquier trayectoria, el trabajo que sobre él realiza el campo gravitatorio de M es $-\Delta E_p = -\big(E_p(y) - E_p(x)\big)$. Si el cuerpo

se mueve sobre una trayectoria γ y sobre él actúa otra fuerza F distinta de la gravitatoria, el trabajo total realizado sobre él es

$$\int_\gamma F\,d\vec{r} - \Delta E_p.$$

Según hemos visto antes, este trabajo es igual al incremento de la energía cinética del cuerpo ΔE_c. Si llamamos *energía total* del cuerpo a $E = E_p + E_c$ concluimos que

$$\Delta E = \Delta E_p + \Delta E_c = \int_\gamma F\,d\vec{r}.$$

En resumen: el trabajo realizado sobre un cuerpo por las fuerzas distintas del campo es igual al incremento de la energía total del cuerpo. En particular, si un cuerpo se mueve sometido únicamente a la acción del campo su energía total permanece constante.

Toda la teoría se desarrolla más cómodamente sin particularizar a un cuerpo m en concreto. Para ello se definen el vector *intensidad de campo* y el *potencial gravitatorio* como

$$E = -\frac{GM}{\|x\|^3}\,x, \quad V = -\frac{GM}{\|x\|},$$

respectivamente, de modo que la fuerza gravitatoria que actúa sobre un cuerpo de masa m es mE y la energía potencial de un cuerpo de masa m es mV. La relación entre ambos es $E = -\nabla V$.

Retomemos los cálculos que hicimos en el capítulo VI sobre un cuerpo que sigue una trayectoria cónica sometido a la fuerza gravitatoria. Puesto que los sumandos de (6.1) son ortogonales deducimos que el cuadrado del módulo de la velocidad es

$$v^2 = \rho'^2 + \rho^2\omega^2.$$

(Notemos que en el capítulo VI llamábamos v al vector velocidad y aquí a su módulo). La energía total del móvil será

$$E = E_c + E_p = \frac{1}{2}m(\rho'^2 + \rho^2\omega^2) - \frac{GMm}{\rho}.$$

Por otro lado la ecuación de la trayectoria es

$$\rho = \frac{L^2}{GMm^2}\frac{1}{1+\epsilon\cos\theta}$$

donde ϵ es la excentricidad de la cónica. Puesto que la energía total es constante, podemos calcularla en el punto que nos resulte más conveniente. Por ejemplo cuando $\theta = 0$, que corresponde con el valor mínimo de ρ, luego $\rho' = 0$. Entonces

$$E_c = \frac{1}{2}m\rho^2\omega^2 = \frac{L^2}{2m\rho^2} = \frac{G^2M^2m^3}{2L^2}\,(1+\epsilon)^2, \quad E_p = -\frac{G^2M^2m^3}{L^2}\,(1+\epsilon),$$

luego

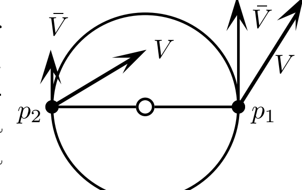
$$E = \frac{G^2M^2m^3}{2L^2}\big((1+\epsilon)^2 - 2(1+\epsilon)\big),$$

y, en definitiva, la energía del móvil es

$$E = \frac{G^2M^2m^3}{2L^2}(\epsilon^2 - 1).$$

Notamos que la trayectoria es elíptica, parabólica o hiperbólica según si $E < 0$, $E = 0$ o $E > 0$. ■

Ejemplo Veamos otra interpretación de la circulación de un campo, ahora en el contexto de la hidrodinámica. Supongamos que V es el campo de velocidades de un fluido. Esto significa que si liberamos una partícula de masa despreciable en un punto p el fluido la arrastrará con velocidad $V(p)$ (no excluimos que V pueda depender del tiempo además de hacerlo de la posición). Supongamos ahora que en el fluido situamos una bolita sujeta por una varilla rígida a un eje, respecto al cual puede girar a lo largo de una circunferencia de radio r.[2] Es claro que si la bolita se encuentra en el punto p el fluido la hará moverse con velocidad igual a la proyección de $V(p)$ sobre la recta tangente a la circunferencia en p, pues la componente normal de la velocidad será cancelada por las fuerzas que mantienen rígida a la varilla que sujeta la bola.



Imaginemos ahora que el eje sujeta a la varilla por el centro y que ésta tiene una bolita en cada brazo. Si éstas se encuentran en los puntos p_1 y p_2, entonces su velocidad (que en módulo ha de ser la misma para ambas a causa de la rigidez de la varilla) estará determinada por los vectores $V(p_1)$ y $V(p_2)$. Al igual que en el caso anterior en realidad dependerá sólo de las proyecciones $\bar{V}(p_1)$ y $\bar{V}(p_2)$ de dichos vectores sobre las rectas tangentes respectivas. Por ejemplo, en el caso indicado en la figura, donde $\|\bar{V}(p_1)\| = 2$ y $\|\bar{V}(p_2)\| = 1$, la velocidad resultante será el promedio[3] de ambas: la varilla girará en sentido contrario a las agujas del reloj con velocidad $(2-1)/2 = 1/2$.

Supongamos ahora que en vez de una varilla tenemos un molinillo con n aspas. Entonces el módulo de la velocidad resultante será

$$\frac{1}{n}V(p_1)T(p_1) + \cdots + \frac{1}{n}V(p_n)T(p_n),$$

donde T es el vector tangente a la circunferencia. Equivalentemente podemos escribir

$$\frac{1}{2\pi r}\left(\frac{2\pi r}{n}V(p_1)T(p_1) + \cdots + \frac{2\pi r}{n}V(p_n)T(p_n)\right),$$

[2]En esta clase de situaciones suponemos siempre que los objetos que introducimos son instrumentos de medida ideales, es decir, que son afectados por el fluido pero ellos no afectan al mismo.

[3]Se trata de un problema de conservación de la cantidad de movimiento. De hecho es equivalente al siguiente: dos cuerpos de la misma masa se aproximan frontalmente de modo que sus velocidades son v_1 y v_2. Si tras el choque se mueven conjuntamente, ¿a qué velocidad lo hacen? La respuesta es que la cantidad de movimiento del sistema es $mv_1 + mv_2$ al principio y $2mv$ al final. Igualando resulta que $v = (v_1 + v_2)/2$. El fluido comunica una cantidad de movimiento a las bolitas y la varilla s limita a unificar las velocidades sin alterar la cantidad de movimiento.

donde r es el radio de la circunferencia. Esto equivale a considerar la circunferencia dividida en n partes iguales de longitud $\Delta s = 2\pi r/n$, multiplicar la longitud de cada parte por el valor de VT en uno de sus puntos, sumar y luego dividir el resultado entre la longitud completa de la circunferencia. Finalmente, si en lugar de un molinillo ponemos una ruedecita de radio r, la velocidad que le imprimirá el fluido vendrá dada por

$$v = \frac{1}{2\pi r}\int_C VT\,ds = \frac{1}{2\pi r}\int_C V\,d\vec{r}.$$

La velocidad v corresponde a una velocidad angular $\omega = v/r$. Así pues,

$$\omega = \frac{1}{2\pi r^2}\int_C V\,d\vec{r}.$$

■

Estudiemos ahora la noción de flujo de un campo a través de una variedad. A cada campo $F : U \subset \mathbb{R}^m \longrightarrow \mathbb{R}^m$, donde U es un abierto en $\mathbb{R}^m$, podemos asociarle la $m-1$-forma

$$d\Phi(F) = \sum_{i=1}^m (-1)^{i+1} F_i\,dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_m.$$

Supongamos que $S \subset U$ es una variedad orientable de dimensión $m-1$, sea n la determinación del vector normal que induce su orientación, tomemos $p \in S$ y $v_1, \ldots, v_{m-1} \in T_p(S)$. Vamos a probar que $d\Phi(F)(p)(v_1, \ldots, v_{m-1})$ es el determinante de la matriz A cuyas filas son $F(p), v_1, \ldots, v_{m-1}$. En efecto, desarrollándolo por la primera fila tenemos que

$$\det A = \sum_{i=1}^m (-1)^{i+1} F_i(p) \det A_i,$$

donde A_i es la matriz que tiene por filas a los vectores $v_1, \ldots, v_{m-1}$ sin su i-ésima componente. Teniendo en cuenta el teorema 9.10 resulta que

$$\det A_i = dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_m(v_1, \ldots v_{m-1})$$

y tenemos la relación buscada. Notemos que para este cálculo no necesitamos que F esté definido más que sobre los puntos de S. En particular podemos aplicarlo al vector normal n, pero entonces el determinante de la matriz cuyas filas son $n(p), v_1, \ldots, v_{m-1}$ es la medida (orientada) del paralelepípedo determinado por estos vectores, y como $n(p)$ es unitario y perpendicular a los restantes, es claro que coincide con la medida orientada de $(v_1, \ldots, v_{m-1})$ en $T_p(S)$. Así pues, si llamamos $d\sigma$ al elemento de medida de S, hemos probado que

$$d\sigma(p) = \sum_{i=1}^m (-1)^{i+1} n_i(p) dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_m.$$

Volviendo al campo F, sobre los puntos de S podemos descomponerlo como $F = (Fn)n+t$, donde $t \in T_p(S)$. Esto nos permite descomponer el determinante de A en dos términos, pero el sumando correspondiente a t es nulo (pues sus filas son m vectores de $T_p(S)$), luego en definitiva

$$d\Phi(F) = (Fn)\left(\sum_{i=1}^{m}(-1)^{i+1}n_i(p)dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_m\right),$$

es decir, $d\Phi(F) = (Fn)\, d\sigma$.

Ejercicio: En particular, si S es una superficie en $\mathbb{R}^3$ tenemos que

$$d\sigma = n_1\, dy \wedge dz + n_2\, dz \wedge dx + n_3\, dx \wedge dy.$$

Calcular a partir de aquí el área de la esfera.

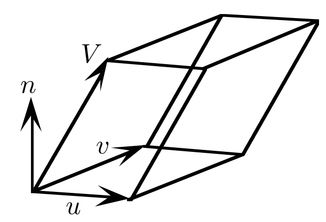
Definición 9.28 Sea $F : U \subset \mathbb{R}^m \longrightarrow \mathbb{R}^m$ un campo de clase C^1 en un abierto $U \subset \mathbb{R}^m$. Se llama *flujo* de F a través de una variedad orientable $S \subset U$ de dimensión $m-1$ a la integral

$$\Phi(F) = \int_S (Fn)\, d\sigma = \int_S \left(\sum_{i=1}^{m}(-1)^{i+1}F_i\, dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_m\right).$$

donde n es el vector normal de S.

En resumen, así como la circulación de un campo a través de una curva era la integral del módulo orientado de la componente tangencial del campo en cada punto de la curva, el flujo de un campo a través de una variedad es la integral del módulo orientado de su componente normal. La hidrodinámica nos proporciona una imagen más concreta del flujo de un campo.

Ejemplo Supongamos que V es la velocidad de un fluido en cada punto. Sea S una superficie en el seno del fluido y $p \in S$. Consideremos un entorno paralelogramo (u, v) en $T_p(S)$ suficientemente pequeño como para que $p + P(u, v)$ se confunda con un subconjunto de S. También podemos suponer que la velocidad del fluido en un entorno de p en $\mathbb{R}^3$ que contiene al paralelogramo es aproximadamente igual a $V(p)$. Nos preguntamos cuál es el volumen de fluido que atraviesa el paralelogramo por unidad de tiempo.



Observamos que el fluido que en un instante dado se encuentra en el paralelogramo, al cabo de una unidad de tiempo se encontrará en otro paralelogramo similar trasladado del primero mediante el vector $V = V(p)$. El volumen de fluido que ha atravesado el paralelogramo será igual al volumen del paralelepípedo (u, v, V). Una simple aplicación del teorema de Fubini muestra que el volumen del paralelepípedo

es igual al área de su base multiplicada por su altura (medida perpendicularmente a la base), es decir, el volumen que atraviesa el paralelogramo es $(Vn)(p)d\sigma(p)(u,v) = d\Phi(V)(u,v)$, entendiendo que el volumen es positivo si el fluido atraviesa el paralelogramo en la dirección de n y negativo en caso contrario.

La aplicación que a cada región de S le asigna el volumen de fluido que lo atraviesa es una medida en S que en un entorno de cada punto debe confundirse con $d\Phi(V)$. Por consiguiente dicho volumen es precisamente $\Phi(V)$.

Si en lugar de hablar de volúmenes queremos hablar de masa habremos de considerar la densidad ρ del fluido en cada punto, es decir, ρ es la derivada de la medida que a cada región del espacio le hace corresponder la masa de fluido que contiene, de modo que dicha masa se recupera integrando ρ en la región en cuestión. Si en lugar de trabajar con V trabajamos con el campo $A = \rho V$, el razonamiento anterior nos da claramente que la cantidad de masa que atraviesa una superficie S es el flujo de A a través de S. ■

Capítulo X

El teorema de Stokes

En el capítulo anterior hemos visto el teorema 9.25, que es la versión para 1-formas de la regla de Barrow. Existe una versión general del teorema de Barrow para n-formas, la cual constituye el último de los resultados fundamentales del cálculo integral de varias variables. Se trata del llamado teorema de Stokes generalizado, del que nos ocuparemos en este capítulo. Aunque todavía no tenemos definidos algunos de los conceptos que involucra, conviene anticipar su aspecto. Se trata de la fórmula:

$$\int_S d\omega = \int_{\partial S} \omega.$$

Aquí ω es una $n-1$-forma definida en una variedad S de dimensión n. En este capítulo extenderemos el concepto de diferencial —que hasta ahora sólo tenemos definido para 0-formas— de modo que la $n-1$-forma ω tendrá asociada una n-forma $d\omega$, que es la que aparece en el primer miembro de la fórmula anterior. También hemos de introducir el concepto de frontera de una variedad S. Por ejemplo, si S es una bola abierta en $\mathbb{R}^3$, entonces ∂S será la esfera del mismo centro y el mismo radio (que es una variedad de una dimensión menos). El teorema de Stokes afirma en este caso que la integral de la forma $d\omega$ sobre la bola puede obtenerse integrando ω sobre la esfera.

Es importante que la noción de frontera de una variedad que vamos a introducir no siempre coincide con la frontera topológica. Por ejemplo, si S es una semiesfera en $\mathbb{R}^3$, todos sus puntos son puntos frontera desde el punto de vista topológico, mientras que su frontera como variedad la formarán los puntos del ecuador, donde la superficie "termina".

10.1 Variedades con frontera

Definición 10.1 Un *semiespacio* en $\mathbb{R}^n$ es un subconjunto de la forma

$$H = \{x \in \mathbb{R}^n \mid u(x) \leq a\},$$

donde $u : \mathbb{R}^n \longrightarrow \mathbb{R}$ es una aplicación lineal y $a \in \mathbb{R}$.

Obviamente H es cerrado en $\mathbb{R}^n$ y su frontera topológica es

$$\partial H = \{x \in \mathbb{R}^n \mid u(x) = a\}.$$

Cuando hablemos de un subespacio abierto U de un semiespacio $H \subset \mathbb{R}^n$ entenderemos que U es abierto respecto a la topología de H, no necesariamente abierto en $\mathbb{R}^n$. Llamaremos *frontera* de U al conjunto $\partial U = U \cap \partial H$. Es claro que ∂U es la intersección de U con su frontera en $\mathbb{R}^n$, por lo que no depende de H. Además el conjunto U es abierto en $\mathbb{R}^n$ si y sólo si $\partial U = \varnothing$. Los puntos de $U \setminus \partial U$ los llamaremos *puntos interiores* de U. Notemos que estos conceptos de interior y frontera no coinciden con los topológicos.

Sea U un subespacio abierto de un semiespacio en $\mathbb{R}^n$. Diremos que una aplicación $f : U \longrightarrow \mathbb{R}^m$ es *diferenciable* (de *clase* C^k, etc.) en un punto $p \in U$ si existe un entorno abierto W de p en $\mathbb{R}^n$ y una aplicación $g : W \longrightarrow \mathbb{R}^m$ diferenciable (de clase C^k, etc.) en el sentido usual y de modo que $g|_{W\cap U} = f$.

Notemos que si $p \notin \partial U$ entonces podemos tomar $W \subset U$ y la condición equivale a que f sea diferenciable (en el sentido usual) en un entorno de p, es decir, a que f sea diferenciable en p en el sentido usual.

Si $f : U \longrightarrow V$ es un difeomorfismo de clase C^1 entre subespacios abiertos de semiespacios de $\mathbb{R}^n$, entonces $f[\partial U] = \partial V$. En efecto, si $p \notin \partial U$ existe $W \subset U$ entorno abierto de p en $\mathbb{R}^n$ tal que $f|_W$ es diferenciable en el sentido usual, luego $f[W] \subset V$ es abierto en $\mathbb{R}^n$, luego $f(p) \notin \partial V$.

Sea $f : U \longrightarrow \mathbb{R}^m$ una aplicación de clase C^1 definida en un subespacio abierto de un semiespacio y $p \in \partial U$. Para $i = 1, 2$ sea $g_i : W_i \longrightarrow \mathbb{R}^m$ una extensión de clase C^1 tal que W_i es abierto en $\mathbb{R}^n$ y $g_i|_{W_i\cap U} = f$. Entonces $dg_1(p) = dg_2(p)$. En efecto, la aplicación $g = g_1 - g_2$ es de clase C^1 en $W = W_1 \cap W_2$ y es nula en todos los puntos del abierto $W \cap (U \setminus \partial U)$. Sus derivadas parciales serán nulas en dicho abierto y por continuidad también lo serán en p. Por consiguiente $dg(p) = 0$ y así $dg_1(p) = dg_2(p)$.

En consecuencia podemos definir $df(p) = dg(p)$, donde g es cualquier extensión de f a un entorno de p en $\mathbb{R}^n$. En particular podemos hablar de las derivadas parciales sucesivas de f en los puntos frontera de su dominio, las cuales están completamente determinadas por f.

A partir de aquí todos los resultados válidos para funciones $f : U \longrightarrow \mathbb{R}^m$ de clase C^k donde U es un abierto en $\mathbb{R}^n$ se extienden trivialmente al caso en que U es un abierto en un semiespacio. Ahora podemos modificar nuestra definición de variedad para admitir puntos frontera:

Definición 10.2 Un conjunto $S \subset \mathbb{R}^m$ es una *variedad diferenciable (con frontera)* de dimensión $n \leq m$ y de clase C^q si para cada punto $p \in S$ existe un entorno V de p, un abierto U en un semiespacio de $\mathbb{R}^n$ y una función $X : U \longrightarrow \mathbb{R}^m$ de clase C^q de modo que el rango de la matriz JX sea igual a n en todo punto y $X : U \longrightarrow S \cap V$ sea un homeomorfismo.

En el resto de la sección la palabra "variedad" hará referencia a variedades con frontera. Veamos en primer lugar que el teorema 5.4 vale también para variedades con frontera, es decir, que si $X : U \longrightarrow S$ es una carta de una variedad S de clase C^q y $u_0 \in U$, entonces existe un entorno G de u_0 en $\mathbb{R}^n$, un entorno V de $X(u_0)$ en $\mathbb{R}^m$ y una aplicación $g : V \longrightarrow G$ de clase C^q tal que $(X|_{G\cap U})^{-1} = g|_{V\cap S}$.

En efecto, tenemos que $JX(u_0)$ tiene rango máximo, luego n de sus filas tienen determinante no nulo. Por simplicidad podemos suponer que son las primeras. Entonces $X(u) = \big(X_1(u), X_2(u)\big)$, donde X_1 tiene las n primeras coordenadas de X y X_2 las restantes, de modo que $JX_1(u_0)$ tiene determinante no nulo. La función X se extiende a una función de clase C^q en un entorno de u_0 en $\mathbb{R}^n$, luego lo mismo le ocurre a X_1. Sea $\bar{X}_1$ una extensión. Por el teorema de inyectividad local y el teorema de la función inversa obtenemos un entorno G de u_0 en $\mathbb{R}^n$ de modo que $W = \bar{X}_1[G]$ es abierto en $\mathbb{R}^n$, $\bar{X}_1|_G : G \longrightarrow W$ es biyectiva y $(\bar{X}_1|_G)^{-1}$ es de clase C^q.

Sea V' un abierto en $\mathbb{R}^m$ tal que $X[U] = V' \cap S$, sea $\pi : \mathbb{R}^m \longrightarrow \mathbb{R}^n$ la proyección en las n primeras componentes, sea $V = \pi^{-1}[W] \cap V'$ y sea $g = \pi \circ (\bar{X}_1|_G)^{-1}$, que es una función de clase C^q. Claramente $X(u_0) \in V$. Si $p \in V \cap S \subset V' \cap S$ entonces $p = X(u)$, para un cierto $u \in U$, además $X_1(u) \in W$, luego $u \in G \cap U$ y así $p \in X[G \cap U]$.

Recíprocamente, si $X(u) \in X[G \cap U]$, entonces $X_1(u) \in W$, luego tenemos $X(u) \in V \cap S$ y

$$g\big(X(u)\big) = (\bar{X}_1|_G)^{-1}\big(X_1(u)\big) = u = (X|_{G\cap U})^{-1}\big(X(u)\big).$$

Así concluimos que $(X|_{G\cap U})^{-1} = g|_{V\cap S}$. ■

Con esto la prueba del teorema 5.5 se adapta fácilmente para probar la versión para variedades con frontera:

Teorema 10.3 *Sea $S \subset \mathbb{R}^m$ una variedad de dimensión n y de clase C^q. Sea $p \in S$ y $X : U \longrightarrow S \cap V$, $Y : U' \longrightarrow S \cap V'$ dos cartas alrededor de p. Sean $V_0 = V \cap V'$, $U_0 = X^{-1}[V_0]$, $U_0' = Y^{-1}[V_0]$. Entonces U_0 y U_0' son abiertos en semiespacios de $\mathbb{R}^n$ y la aplicación $X \circ Y^{-1} : U_0 \longrightarrow U_0'$ es biyectiva, de clase C^q y con determinante jacobiano no nulo, con lo que su inversa es también de clase C^q.*

Definición 10.4 Sea S una variedad con frontera. Un punto $p \in S$ es un *punto frontera* de S si cuando $X : U \longrightarrow S$ es una carta alrededor de p entonces las coordenadas $X^{-1}(p)$ son un punto frontera de U. Por el teorema anterior esto no depende de la elección de la carta. Llamaremos *frontera* de S al conjunto ∂S de puntos frontera de S.

Es claro que ∂S es cerrado en S (quizá vacío) y que $S \setminus \partial S$ es una variedad en el sentido del capítulo V. De este modo las variedades sin frontera coinciden con las variedades con frontera cuya frontera es vacía.

A partir de aquí toda la teoría de variedades diferenciales se generaliza sin dificultad para el caso de variedades con frontera: podemos definir los espacios

tangentes, la diferenciabilidad y la diferencial de aplicaciones entre variedades, etc. todo sin cambio alguno. Sólo hay una salvedad: podemos definir el producto de una variedad con frontera por una variedad sin frontera, pero no el producto de dos variedades con frontera. La razón es que el producto de un abierto en un semiespacio de $\mathbb{R}^m$ por un abierto en $\mathbb{R}^n$ es un abierto en un semiespacio en $\mathbb{R}^{m+n}$, mientras que el producto de dos abiertos en dos semiespacios no es necesariamente un abierto en un semiespacio. Por ejemplo, el producto de dos semirrectas cerradas es un cuadrante cerrado en $\mathbb{R}^2$, que no es un abierto en ningún semiplano. Para admitir productos de variedades con frontera tendríamos que definir variedades con esquinas.

Por otra parte no es necesario generalizar el cálculo integral al caso de variedades con frontera. Si el lector se molesta en hacerlo demostrará que la frontera de una variedad siempre tiene medida nula, luego en la práctica podemos integrar en el conjunto de puntos interiores.

Ejemplo Una bola cerrada en $\mathbb{R}^n$ es una variedad cuya frontera coincide con su frontera topológica. En efecto, consideremos por ejemplo la bola B de centro 0 y radio 1. Para cubrir los puntos interiores tomamos como carta la identidad en la bola abierta. Veamos ahora una carta que cubre los puntos frontera de la semiesfera $x_n > 0$. Similarmente se cubren los puntos restantes. Definimos $X : \mathbb{R}^{n-1} \times]0, 1] \longrightarrow B$ mediante

$$X(u, r) = \frac{(ru, r)}{\|(u, 1)\|}.$$

Claramente el dominio $U = \mathbb{R}^{n-1} \times]0, 1]$ es un abierto en el semiplano $x_n \leq 1$ y X puede extenderse hasta el abierto $\mathbb{R}^{n-1} \times]0, +\infty[$ mediante la misma expresión, y ciertamente es de clase C^∞.

Podemos ver X como sigue: dados (u, r) construimos el punto $(u, 1)$ que está en el plano tangente a la esfera por su polo norte, dividimos por su norma, con lo que pasamos a un punto de la esfera (distinto para cada valor de u) y luego multiplicamos por r, con lo que pasamos a un punto de la esfera de radio r. Teniendo esto en cuenta es claro que X es inyectiva. Concretamente su inversa es

$$X^{-1}(x) = \left(\frac{x_1}{x_n}, \ldots, \frac{x_{n-1}}{x_n}, \|x\|\right),$$

que es diferenciable, por lo que las diferenciales de X y X^{-1} son mutuamente inversas, luego JX tiene rango n. ■

Ejercicio: Probar que una corona esférica cerrada (es decir, una bola cerrada menos una bola abierta concéntrica) es una variedad con frontera.

Ejercicio: Probar que un casquete esférico cerrado es una variedad con frontera.

Ejercicio: Probar que un cuadrado cerrado menos sus cuatro vértices es una variedad con frontera. (Con los vértices sería una variedad con esquinas.)

Ejercicio: Definimos el *toro sólido* de radios r y R (tales que $0 < r < R$) como la imagen de $\mathbb{R}^2 \times [0, r]$ por la aplicación

$$X(u, v, \rho) = (R \cos v + \rho \cos u \cos v, R \operatorname{sen} v + \rho \cos u \operatorname{sen} v, \rho \operatorname{sen} u).$$

Probar que es una variedad cuya frontera es igual al toro de radios r y R.

Vamos a probar que la frontera de una variedad S es a su vez una variedad (sin frontera) de una dimensión menos. Para ello probamos el teorema siguiente, que además nos permitirá relacionar las orientaciones de S y ∂S.

Teorema 10.5 *Sea S una variedad de dimensión $n > 1$ y sea $p \in \partial S$. Entonces p puede cubrirse por una carta $X : U \longrightarrow V \cap S$ tal que $U =]-1, 0] \times]-1, 1[^{n-1}$, los puntos de $V \cap \partial S$ son exactamente los puntos de $V \cap S$ tales que $x_1 = 0$, los puntos de $V \cap S \setminus \partial S$ son los puntos de $V \cap S$ tales que $x_1 < 0$ y $p = X(0, \ldots, 0)$. Si S es orientable la carta puede tomarse positiva.*

DEMOSTRACIÓN: En principio podemos tomar una carta cuyo dominio sea un abierto en un semiespacio de modo que los puntos de la frontera tengan coordenadas en un cierto hiperplano afín de $\mathbb{R}^n$. Componiendo la carta con una aplicación afín (que es de clase C^∞) podemos exigir que dicho hiperplano sea $x_1 = 0$ y que el semiespacio de coordenadas sea $x_1 \leq 0$. Más aún, podemos exigir que las coordenadas de p sean nulas. El dominio de la carta será un entorno de 0, luego contendrá un semicubo de centro 0. Componiendo con una homotecia podremos conseguir que contenga el semicubo U del enunciado y restringiendo la carta a U tenemos una carta en las condiciones pedidas. Si S está orientada y la carta que hemos obtenido es negativa basta tomar la carta $\bar{X}(x_1, \ldots, x_n) = X(x_1, \ldots, x_{n-1}, -x_n)$ y pasamos a tener una carta positiva (aquí usamos que $n > 1$). ■

Teorema 10.6 *Sea S una variedad de dimensión n y clase C^q con frontera no vacía. Entonces ∂S es una variedad (sin frontera) de dimensión $n - 1$ y clase C^q. Basta tomar como cartas las de la forma $Y(x_2, \ldots, x_n) = X(0, x_2, \ldots, x_n)$, donde X es una carta del tipo dado por el teorema anterior (con lo que Y está definida sobre el cubo $]-1, 1[^{n-1}$). Si S es orientable y tomamos sólo cartas positivas obtenemos un atlas orientado para ∂S.*

DEMOSTRACIÓN: Por el teorema anterior todo punto de ∂S es cubrible por una de estas cartas. Claramente son de clase C^q y $JY(x_2, \ldots, x_n)$ se obtiene quitando la primera fila a $JX(0, x_2, \ldots, x_n)$, luego su rango es $n - 1$. También es evidente que $Y :]-1, 1[^{n-1} \longrightarrow V \cap \partial S$ es un homeomorfismo.

Supongamos ahora que S es orientable y que tenemos dos cartas positivas Y_1 e Y_2 que cubran a un mismo punto, obtenidas a partir de cartas X_1 y X_2. Consideremos la aplicación $g = X_1^{-1} \circ X_2$. Tenemos que $g_1(0, x_2, \ldots, x_n) = 0$ para todas las coordenadas $(x_2, \ldots, x_n)$, pues g hace corresponder puntos frontera de U_1 con puntos frontera de U_2. Por lo tanto $D_i g_1(x_2, \ldots, x_n) = 0$

para todo $i > 1$, luego la matriz jacobiana de g es de la forma

$$Jg(0, x_2, \ldots, x_n) = \left(\begin{array}{c|ccc} D_1 g_1 & * & \cdots & * \\ \hline 0 & & & \\ \vdots & & A & \\ 0 & & & \end{array}\right)$$

Sea $h = Y_1^{-1} \circ Y_2$. Claramente $h = \iota \circ g \circ \pi$, donde $\iota(x_2, \ldots, x_n) = (0, x_2, \ldots, x_n)$ y $\pi(x_1, \ldots, x_n) = (x_2, \ldots, x_n)$. Considerando las matrices jacobianas concluimos fácilmente que $Jh(x_2, \ldots, x_n) = A$. Así pues

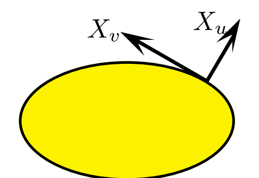
$$|Jg(0, x_2, \ldots, x_n)| = D_1 g_1(0, x_2, \ldots, x_n)\,|Jh(x_2, \ldots, x_n)|.$$

Como las cartas X_1 y X_2 son positivas tenemos que primer determinante es positivo. Si probamos que $D_1 g_1(0, x_2, \ldots, x_n) > 0$ tendremos que el determinante de Jh será positivo también, y esto probará que Y_1 e Y_2 tienen la misma orientación. Ahora bien,

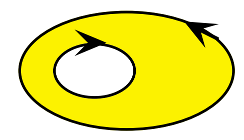
$$D_1 g_1(0, x_2, \ldots, x_n) = \lim_{r \to 0} \frac{g_1(r, x_2, \ldots, x_n)}{r}.$$

Si $r < 0$ estamos en las coordenadas de un punto de S, luego $g(r, x_2, \ldots, x_n)$ es la primera coordenada de dicho punto en la carta X_2, luego $g(r, x_2, \ldots, x_n) < 0$, luego el cociente es positivo y el límite (que ciertamente es no nulo) es también positivo. ■

En lo sucesivo, cuando S sea una variedad orientada sobreentenderemos que ∂S tiene la orientación dada por el teorema anterior.



Ejemplo Consideremos la variedad con frontera S indicada en la figura dotada de la orientación usual. Entonces ∂S queda orientada de modo que al recorrerla en sentido positivo giramos en sentido contrario a las agujas del reloj. En efecto, dado un punto p en la frontera de coordenadas $(0, v_0)$, sea X una carta positiva que lo cubra. La curva $X(u, v_0)$ está dentro de S cuando $u < 0$ y fuera de S cuando $u > 0$ (recordemos que las coordenadas se pueden extender un poco fuera de S). Por consiguiente el vector X_u apuntará hacia fuera de S. Como la base (X_u, X_v) ha de ser positiva, el vector X_v ha de marcar la dirección de giro antihoraria. En el teorema anterior hemos adoptado los convenios necesarios para que esto sea así.



En cambio, si la variedad tiene un agujero, la frontera del mismo quedará orientada en sentido horario (el vector X_u apuntará hacia el interior del agujero). Es fácil ver en general que si $V \subset \mathbb{R}^m$ es una variedad de dimensión m con la orientación natural en $\mathbb{R}^m$ (la dada por la base canónica), entonces la orientación en ∂V es la inducida por el vector normal que apunta hacia fuera de V.

10.2 La diferencial exterior

Recordemos que nos dirigimos hacia la prueba del teorema de Stokes generalizado, que nos da una relación de la forma

$$\int_S d\omega = \int_{\partial S} \omega.$$

Ahora ya sabemos qué debemos entender por ∂S, pero nos falta definir la diferencial de una k-forma arbitraria ω, que ha de ser una $k+1$-forma. La idea básica es muy simple: la diferencial de la n-forma $f\,dx_1 \wedge \cdots \wedge dx_k$ será la $k+1$-forma $df \wedge dx_1 \wedge \cdots \wedge dx_k$. Sin embargo, esto no nos sirve como definición, pues hemos de justificar que el resultado no depende del sistema de coordenadas con el que trabajamos. Como al diferencial perdemos un grado de derivabilidad, para evitar cuestiones técnicas al respecto a partir de ahora supondremos que todas las variedades y funciones que consideremos serán de clase C^∞. El lector puede entretenerse relajando las hipótesis de derivabilidad en cada caso hasta las estrictamente necesarias. Todos los resultados valen para variedades con frontera.

Teorema 10.7 *Si S es una variedad diferenciable existe una única aplicación lineal $d : \Lambda(S) \longrightarrow \Lambda(S)$ que cumple las propiedades siguientes:*

a) Si $f \in \Lambda^0(S)$, entonces df es la diferencial de f en el sentido usual.

b) Para cada $\omega \in \Lambda^k(S)$ se cumple que $d\omega \in \Lambda^{k+1}(S)$.

c) Si $\omega_1 \in \Lambda^k(S)$ y $\omega_2 \in \Lambda(S)$, entonces

$$d(\omega_1 \wedge \omega_2) = d\omega_1 \wedge \omega_2 + (-1)^k \omega_1 \wedge d\omega_2.$$

d) $d^2 = d \circ d = 0$.

e) Si $\omega \in \Lambda(S)$ se anula en un abierto $V \subset S$ entonces $d\omega$ también se anula en V.

Demostración: En primer lugar probaremos que la última propiedad es consecuencia de las anteriores. Para cada $p \in V$ existe una función $\{p\} \prec f \prec V$. Entonces la forma $f\omega$ es nula, y como la diferencial es lineal ha de ser

$$0 = d(f\omega) = df \wedge \omega + f \wedge d\omega,$$

luego $d\omega(p) = (f \wedge d\omega)(p) = -df(p) \wedge 0 = 0$.

Notemos que la propiedad e) junto con la linealidad de la diferencial prueba que si dos formas coinciden en un abierto de S entonces sus diferenciales también coinciden.

Ahora probamos que si existe la diferencial es única. Tomemos un punto $p \in S$ y sea $X : U \longrightarrow V \subset S$ una carta alrededor de p. Tomemos un entorno de p cuya clausura K sea compacta y esté contenida en V. Sea $K \prec f \prec V$.

Si $\omega \in \Lambda^k(S)$ entonces $\omega|_V$ se expresa como

$$\sum_{1\leq i_1<\cdots<i_k\leq n} \omega_{i_1\cdots i_k} dx_{i_1} \wedge \cdots \wedge dx_{i_k}, \tag{10.1}$$

para ciertas funciones $\omega_{i_1\cdots i_k}$ de clase C^∞ en V.

La forma $f\omega$ coincide con ω en un entorno de p y sus coeficientes son las funciones $\bar{\omega}_{i_1\cdots i_k} = f\omega_{i_1\cdots i_k}$. Estas funciones se anulan fuera de un compacto contenido en V, luego podemos extenderlas a funciones de clase C^∞ en S haciendo que tomen el valor 0 fuera de V. Similarmente, las funciones $y_i = fx_i$ extendidas como 0 fuera de U son de clase C^∞ en S y coinciden con las x_i en un entorno de p. Por consiguiente dy_i coincide con dx_i en un entorno de p. Así pues, la forma $f\omega$ (y por consiguiente ω) coincide con la forma

$$\bar{\omega} = \sum_{1\leq i_1<\cdots<i_k\leq n} \bar{\omega}_{i_1\cdots i_k} dy_{i_1} \wedge \cdots \wedge dy_{i_k}$$

en un entorno de p. Ahora calculamos

$$d\bar{\omega} = \sum_{1\leq i_1<\cdots<i_k\leq n} d\bar{\omega}_{i_1\cdots i_k} \wedge dy_{i_1} \wedge \cdots \wedge dy_{i_k},$$

pues una simple inducción prueba a partir de c) y d) que $d(dy_{i_1} \wedge \cdots \wedge dy_{i_k}) = 0$. Teniendo en cuenta que la diferencial depende sólo del comportamiento local de las formas llegamos a que

$$d\omega(p) = \sum_{1\leq i_1<\cdots<i_k\leq n} d\omega_{i_1\cdots i_k}(p) \wedge dx_{i_1}(p) \wedge \cdots \wedge dx_{i_k}(p). \tag{10.2}$$

Ahora bien, por la propiedad a) tenemos que el miembro derecho de la igualdad anterior es el mismo cualquiera que sea la función d que cumpla las propiedades del enunciado. En consecuencia la diferencial exterior es única.

Veamos que toda variedad cubrible por una sola carta tiene una diferencial exterior. En tal caso toda k-forma ω se expresa de forma única como (10.1). Para cada punto p definimos $d\omega(p)$ mediante (10.2). Claramente $d\omega$ así definida es una $k+1$-forma (de clase C^∞) y la diferencial d es lineal. Las únicas propiedades que no son evidentes son c) y d).

Puesto que los dos miembros de la igualdad c) son lineales en ω_1 y ω_2, basta probar que se da la igualdad cuando

$$\omega_1 = f_1\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}, \quad \omega_2 = f_2\, dx_{j_1} \wedge \cdots \wedge dx_{j_r}.$$

Entonces

$$\omega_1 \wedge \omega_2 = f_1 f_2\, dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_r},$$

luego

$$d(\omega_1 \wedge \omega_2) \quad = \quad (f_2\, df_1 + f_1\, df_2) \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_r}$$

$$\begin{aligned}
&= f_2\, df_1 \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_r} \\
&\quad + f_1\, df_2 \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_r} \\
&= df_1 \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge f_2\, dx_{j_1} \wedge \cdots \wedge dx_{j_r} \\
&\quad + (-1)^k f_1 dx_{i_1} \wedge \cdots \wedge dx_{i_k} \wedge df_2 \wedge dx_{j_1} \wedge \cdots \wedge dx_{j_r} \\
&\quad d\omega_1 \wedge \omega_2 + (-1)^k \omega_1 \wedge d\omega_2.
\end{aligned}$$

Del mismo modo, basta comprobar que d) se cumple para una forma de tipo

$$\omega = f\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

Entonces

$$d\omega = df \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k}$$

y basta ver que $d^2 = 0$ al actuar sobre 0-formas, pues una inducción usando la propiedad c) ya demostrada nos da el caso general de d). Veamos, pues, que $d(df)) = 0$. En efecto, sabemos que

$$df = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}\, dx_i.$$

En consecuencia

$$\begin{aligned}
d(d(f)) &= \sum_{i=1}^{n} d\frac{\partial f}{\partial x_i} \wedge dx_i = \sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f}{\partial x_i \partial x_j} dx_j \wedge dx_i \\
&= \sum_{1 \le i < j \le n} \left(\frac{\partial^2 f}{\partial x_i \partial x_j} - \frac{\partial^2 f}{\partial x_j \partial x_i} \right) dx_j \wedge dx_i = 0.
\end{aligned}$$

Consideremos finalmente una variedad arbitraria S. Si $p \in S$ y V es la imagen de una carta que cubre a p, entonces tenemos definida una diferencial exterior d_V en $\Lambda(V)$. Para cada forma $\omega \in \lambda(S)$, definimos $d\omega(p) = d_V(\omega|_V)(p)$. Veamos que la forma $d\omega(p)$ no depende de la elección de V.

Si tomamos dos cartas X e Y con imágenes V_1 y V_2 entonces $V_1 \cap V_2$ también es la imagen de una carta, luego también tenemos definida una diferencial exterior en $\Lambda(V_1 \cap V_2)$. Ahora bien, si $\omega|_V$ admite la expresión (10.1), entonces $\omega|_{V_1 \cap V_2}$ admite la misma expresión (restringiendo las funciones $\omega_{i_1 \cdots i_k}$ y x_{i_j} a $V_1 \cap V_2$). Por consiguiente $d_V(\omega|_{V_1})(p) = d_{V_1 \cap V_2}(\omega|_{V_1 \cap V_2})(p)$, pues ambas vienen dadas por (10.2). Por otro lado $d_V(\omega|_{V_2})(p) = d_{V_1 \cap V_2}(\omega|_{V_1 \cap V_2})(p)$. Ahora usamos que la diferencial en $\Lambda(V_1 \cap V_2)$ es la misma independientemente de la carta con la que se calcula (por la unicidad que hemos probado) y concluimos que $d_V(\omega|_{V_1})(p) = d_V(\omega|_{V_2})(p)$, como queríamos probar.

Con esto tenemos definida una aplicación $d : \Lambda(S) \longrightarrow \Lambda(S)$. Notemos que $d\omega$ es realmente una forma de clase C^∞ porque si $p \in S$ y V es la imagen de una carta que cubre a p tenemos que $(d\omega)|_V = d_V(\omega|_V)$ y ésta es una forma de clase C^∞. Esta misma relación justifica también que d es lineal, así como que verifica las propiedades del enunciado. ■

Si no suponemos que las formas son de clase C^∞ podemos definir igualmente la diferencial de una forma de clase C^k como una forma de clase C^{k-1} y todas las propiedades del teorema anterior se cumplen igualmente con las restricciones obvias. Por ejemplo, para $d(d\omega)) = 0$ hemos de exigir que ω sea al menos de clase C^2. Veamos algunos casos particulares de la diferencial exterior:

Definición 10.8 Sea $F : U \subset \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ un campo de vectores definido sobre un abierto U. El *rotacional* de F es el campo $\operatorname{rot} F : U \longrightarrow \mathbb{R}^3$ dado por

$$\operatorname{rot} F = \begin{vmatrix} e_1 & e_2 & e_3 \\ \dfrac{\partial}{\partial x} & \dfrac{\partial}{\partial y} & \dfrac{\partial}{\partial z} \\ F_1 & F_2 & F_3 \end{vmatrix} = \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z}, \ \frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x}, \ \frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right).$$

Por supuesto que el determinante intermedio es sólo una regla mnemotécnica, que también puede abreviarse como $\operatorname{rot} F = \nabla \wedge F$. La relación con la diferencial exterior es la siguiente: según vimos en el capítulo anterior, al campo F le podemos asociar la 1-forma $F\,d\vec{r} = F_1\,dx + F_2\,dy + F_3\,dz$. La diferencial de esta forma es

$$\begin{aligned} d(F\,d\vec{r}) &= dF_1 \wedge dx + dF_2 \wedge dy + dF_3 \wedge dz = \left(\frac{\partial F_3}{\partial y} - \frac{\partial F_2}{\partial z} \right) dy \wedge dz \\ &+ \left(\frac{\partial F_1}{\partial z} - \frac{\partial F_3}{\partial x} \right) dz \wedge dx + \left(\frac{\partial F_2}{\partial x} - \frac{\partial F_1}{\partial y} \right) dy \wedge dz \\ &= d\Phi(\operatorname{rot} F)). \end{aligned}$$

Es decir, la diferencial del elemento de circulación de F es el elemento de flujo del rotacional de F.

Si $f : U \longrightarrow \mathbb{R}$ es un campo escalar, es claro que $df = \nabla f\,d\vec{r}$, luego

$$0 = d^2 f = d(\nabla f\,d\vec{r}) = d\Phi(\operatorname{rot} \nabla f)),$$

de donde se sigue la relación

$$\operatorname{rot} \nabla f = 0,$$

para todo campo escalar f.

Definición 10.9 Sea $F : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ un campo de clase C^∞ en un abierto U. La *divergencia* de F es el campo escalar $\operatorname{div} F : U \longrightarrow \mathbb{R}$ dado por

$$\operatorname{div} F = \frac{\partial F_1}{\partial x_1} + \cdots + \frac{\partial F_n}{\partial x_n}.$$

Claramente

$$\begin{aligned} d(d\Phi(F)) &= \sum_{i=1}^n (-1)^{i+1} dF_i \wedge dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_n \\ &= \sum_{i=1}^n (-1)^{i+1} \frac{\partial F_i}{\partial x_i}\, dx_i \wedge dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_n \\ &= \sum_{i=1}^n \frac{\partial F_i}{\partial x_i}\, dx_1 \wedge \cdots \wedge dx_n = \operatorname{div} F\,dm. \end{aligned}$$

Así pues, la diferencial del elemento de flujo de F es la divergencia por el elemento de volumen de $\mathbb{R}^n$. En particular, si $n = 3$ y aplicamos esto a $\operatorname{rot} F$ resulta

$$\operatorname{div} \operatorname{rot} F\, dm = d(d\Phi(\operatorname{rot} F)) = d(d(F\, d\vec{r})) = 0,$$

luego

$$\operatorname{div} \operatorname{rot} F = 0,$$

para todo campo vectorial F.

El teorema de Stokes, que probaremos en la sección siguiente, no sólo nos permitirá aprovechar estos conceptos que acabamos de introducir, sino que nos dará interpretaciones geométricas de los mismos.

Conviene observar que en algunos contextos usamos la letra d para indicar algo que no es una diferencial exterior. Así por ejemplo, el elemento de medida dm no es la diferencial exterior de ninguna forma m, ni tampoco lo es en general el elemento de flujo de un campo $d\Phi(F)$. En estos caso la d sólo indica que la medida se obtiene integrando dm o que el flujo se obtiene integrando $d\Phi(F)$, pero sería incorrecto afirmar cosas como que $d(d\Phi(F)) = 0$.

Terminamos la sección con una propiedad importante de la diferencial:

Teorema 10.10 *Sea $f : S \longrightarrow T$ una aplicación de clase C^∞ entre variedades. Entonces la retracción $f^\sharp : \Lambda(T) \longrightarrow \Lambda(S)$ conmuta con la diferencial exterior, es decir, $f^\sharp(d\omega) = df^\sharp(\omega)$, para toda $\omega \in \Lambda(T)$.*

Demostración: Es claro que el valor de ambas formas en un punto p depende únicamente de los valores que toma ω en un entorno de $f(p)$, luego no perdemos generalidad si suponemos que T es el rango de una carta Y de coordenadas $y_1, \ldots, y_n$. En tal caso basta probar que

$$f^\sharp\big(d(g\, dy_{i_1} \wedge \cdots \wedge dy_{i_k})\big) = d\big(f^\sharp(g\, dy_{i_1} \wedge \cdots \wedge dy_{i_k})\big).$$

Ahora bien, ambos miembros son $d(f \circ g) \wedge d(f \circ y_{i_1}) \wedge \cdots \wedge d(f \circ y_{i_k})$. ■

10.3 El teorema de Stokes

Ahora ya tenemos todos los elementos necesarios para enunciar el Teorema de Stokes. Sin embargo debemos introducir algunos más que nos servirán de ayuda en la demostración.

Definición 10.11 Un *n-cubo* es un conjunto de la forma

$$S = [a_1, b_1] \times \cdots \times [a_n, b_n],$$

donde $a_i < b_i$ son números reales. La frontera (topológica) de S en $\mathbb{R}^n$ está formada por la unión de los $2n$ conjuntos

$$\begin{aligned} S_i^0 &= [a_1, b_1] \times \cdots \times \{a_i\} \times \cdots \times [a_n, b_n], \\ S_i^1 &= [a_1, b_1] \times \cdots \times \{b_i\} \times \cdots \times [a_n, b_n], \quad i = 1, \ldots, n, \end{aligned}$$

a los que llamaremos *caras* del cubo.

Una forma diferencial en un cubo S es simplemente una forma diferencial definida en un abierto de $\mathbb{R}^n$ que contenga a S. Consideraremos a dicho abierto como variedad orientada, tomando a la identidad como carta positiva (con lo que la base canónica de $\mathbb{R}^n$ es positiva). En particular tenemos definida la integral de una n-forma sobre un n-cubo.

Si ω es una $n-1$-forma en un cubo S, donde $n > 1$, vamos a definir la integral de ω sobre ∂S. Para ello comenzamos definiendo la integral sobre cada cara. Consideramos primero una forma de tipo

$$\omega(x_1, \ldots, x_n) = f(x_1, \ldots, x_n)\, dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_n.$$

La integral de ω sobre la cara S_j^k (para $j = 1, \ldots, n$ y $k = 0, 1$) se define como igual a 0 si $j \neq i$ y en caso contrario mediante

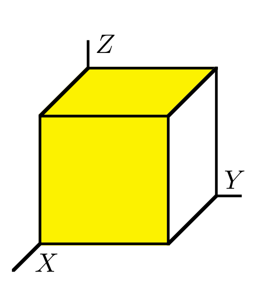
$$\begin{aligned}\int_{S_i^0} \omega &= \int_C f(x_1, \ldots, a_i, \ldots, x_n)\, dx_1 \cdots dx_{i-1}\, dx_{i+1} \cdots dx_n, \\ \int_{S_i^1} \omega &= \int_C f(x_1, \ldots, b_i, \ldots, x_n)\, dx_1 \cdots dx_{i-1}\, dx_{i+1} \cdots dx_n,\end{aligned}$$

donde $C = [a_1, b_1] \times \cdots \times [a_{i-1}, b_{i-1}] \times [a_{i+1}, b_{i+1}] \times \cdots \times [a_n, b_n]$.

Una $n-1$-forma arbitraria se descompone de forma única en suma de n formas del tipo anterior (en cada una de las cuales falta un dx_i distinto). Definimos su integral sobre la cara S_i^k como la suma de las integrales de estas n formas. Así tenemos definida $\int_{S_i^k} \omega$ para cualquier $n-1$-forma sobre S. La integral es obviamente lineal en ω.

Finalmente definimos

$$\int_{\partial S} \omega = \sum_{i=1}^{n} (-1)^i \left(\int_{S_i^0} \omega - \int_{S_i^1} \omega \right).$$



Conviene entender por qué es ésta la definición correcta de integral sobre ∂S. Pensemos por ejemplo en un cubo tridimensional. Según la fórmula las integrales sobre caras opuestas se suman con signos opuestos. Concretamente tienen signo positivo las dos que en la figura aparecen sombreadas más la situada sobre el plano XZ, que no se ve. Nuestra intención es tratar al cubo como si fuese una variedad con frontera. No lo es a causa de que la frontera tiene aristas donde no es diferenciable, pero a efectos de la integración esto no va a afectar porque las aristas tienen área nula, y el teorema de Stokes va a ser cierto también sobre el cubo. La orientación que debemos imponer a la frontera en analogía con las variedades es la inducida por el vector normal que apunta hacia fuera del cubo. Supongamos que queremos integrar una forma de tipo $f(x, y, z)\, dx \wedge dz$. Es claro que sólo van a influir las caras con y constante, pues dx es nula en las caras con x constante y dz es nula en las caras con z constante.

Para integrar la forma sobre S_y^1 (la cara que en la figura queda a la derecha) consideramos la carta $X(x, z) = (x, y_0, z)$. La base asociada en el plano tangente

de cada punto es $X_x = (1,0,0)$, $X_z = (0,0,1)$, y por consiguiente $X_x \wedge X_z = (0,-1,0)$, que apunta hacia dentro del cubo, luego la carta es negativa y la integral es

$$\int_{S^1_y} f\,dx \wedge dz = -\int_C f(x, y_0, z)\,dxdz,$$

y el signo corresponde con el que hemos establecido en la definición. En cambio, si la integral es sobre la cara opuesta, ahora el vector $(0,-1,0)$ sí que apunta hacia fuera del cubo, luego la carta es positiva y no hay que cambiar el signo, tal y como indica la definición.

Mediante este tipo de razonamientos es posible justificar que la definición que hemos dado hace que la integral sobre ∂S sea la correcta respecto a la orientación de las caras inducida por la orientación usual del interior del cubo, es decir, la que hace positiva una base de una cara si al añadirle *como primer vector* uno que apunte hacia fuera del cubo obtenemos una base positiva de $\mathbb{R}^n$. De todos modos esto no es muy importante, pues sólo vamos a usar las integrales sobre cubos como un paso previo a la prueba del teorema de Stokes sobre variedades orientadas.

Ejercicio: Representar gráficamente la orientación de la frontera de un cuadrado según la definición que hemos dado.

Si definimos la integral de una 0-forma sobre la frontera de un 1-cubo, es decir, de una función f sobre los extremos de un intervalo $S = [a, b]$, como

$$\int_{\partial S} f = f(b) - f(a),$$

el teorema siguiente tiene sentido para $n = 1$ y entonces no es más que la regla de Barrow:

Teorema 10.12 (Teorema de Stokes para un cubo) *Sea S un n-cubo y ω una $n-1$-forma en S. Entonces*

$$\int_S d\omega = \int_{\partial S} \omega.$$

DEMOSTRACIÓN: Según acabamos de comentar, el caso $n = 1$ es simplemente la regla de Barrow. Supongamos, pues $n > 1$. Por la linealidad de la integral y de la diferencial es suficiente probar el teorema cuando la forma es

$$\omega(x_1,\ldots,x_n) = f(x_1,\ldots,x_n)\,dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_n.$$

Por definición la integral de ω es nula sobre todas las caras de S excepto S^k_i, para $k = 0, 1$. Así pues, $\int_{\partial S} \omega$ es igual a

$$(-1)^i \int_C \big(f(x_1,\ldots,a_i,\ldots,x_n) - f(x_1,\ldots,b_i,\ldots,x_n)\big)\,dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_n,$$

donde C es el cubo que resulta de eliminar el i-ésimo intervalo a S. Por otro lado,

$$\begin{aligned} d\omega &= df \wedge dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_n \\ &= \frac{\partial f}{\partial x_i} dx_i \wedge dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_n \\ &= (-1)^{i-1} \frac{\partial f}{\partial x_i} dx_1 \wedge \cdots \wedge dx_n. \end{aligned}$$

Así pues,

$$\int_S d\omega = (-1)^{i-1} \int_S \frac{\partial f}{\partial x_i} dx_1 \cdots dx_n.$$

Por el teorema de Fubini podemos integrar primero respecto a dx_i, para lo cual aplicamos la regla de Barrow y queda

$$(-1)^{i-1}\int_C \big(f(x_1,\ldots,b_i,\ldots,x_n)-f(x_1,\ldots,a_i,\ldots,x_n)\big)dx_1\cdots dx_{i-1}dx_{i+1}\cdots dx_n,$$

que coincide con la integral sobre la frontera. ■

Teorema 10.13 (Teorema de Stokes generalizado) *Sea S una variedad diferenciable orientable de dimensión $n > 1$ y ω una $n-1$-forma en S de soporte compacto.*[1] *Entonces*

$$\int_S d\omega = \int_{\partial S} \omega.$$

Demostración: Vamos a definir para cada punto $p \in S$ un entorno V_p en S. Si p es un punto interior tomamos una carta positiva $X_p : U_p \longrightarrow S$ alrededor de p, donde U_p es una bola abierta en $\mathbb{R}^n$. Tomamos un cubo $C_p \subset U_p$ que contenga en su interior al vector de coordenadas de p y llamamos V_p a la imagen por X_p del interior de C_p (que obviamente es un entorno de p en S.

Si $p \in \partial S$ tomamos una carta positiva $X_p : U_p \longrightarrow S$ de modo que

$$U_p = \left]-2, 0\right] \times \left]-2, 2\right[^{n-1},$$

$X_p(0,\ldots,0) = p$ y los puntos de ∂S en $X_p[U_p]$ sean exactamente los que cumplan $x_1 = 0$ (una leve modificación de la prueba del teorema 10.5 prueba la existencia de tal carta). Es claro que X_p induce una carta positiva en ∂S. Llamamos $C_p = [-1, 0] \times [-1, 1]^{n-1}$, que es un entorno de 0 en U_p y tomamos como V_p la imagen por X_p del interior de C_p, es decir, de $\left]-1, 0\right] \times \left]-1, 1\right[^{n-1}$, que es un entorno de p en S.

Los abiertos V_p cubren el soporte de ω, que por hipótesis es compacto, luego existe un subcubrimiento finito formado por los abiertos $V_{p_1}, \ldots, V_{p_r}$. Por el teorema 7.27 existen funciones $h_1, \ldots, h_r$ en S (que según hemos comentado las podemos tomar de clase C^∞) de modo que $h_i \prec V_{p_i}$ y $h_1 + \cdots + h_r$ toma el valor 1 sobre cada punto del soporte de ω.

[1] Un análisis de la prueba muestra que basta exigir que S sea de clase C^2 y ω de clase C^1.

Sea $\omega_i = h_i\omega$, que claramente es una $n-1$-forma (de clase C^∞) con soporte compacto contenido en V_{p_i}. El complementario del soporte de ω_i es un abierto donde ω_i se anula, luego lo mismo le ocurre a $d\omega_i$, es decir, que $d\omega_i$ también tiene el soporte contenido en V_{p_i}. Además $\omega = \omega_1 + \cdots + \omega_r$, luego

$$\int_S d\omega = \sum_{i=1}^r \int_S d\omega_i = \sum_{i=1}^r \int_{V_{p_i}} d\omega_i, \qquad \int_{\partial S} \omega = \sum_{i=1}^r \int_{\partial S} \omega_i = \sum_{i=1}^r \int_{V_{p_i}\cap\partial S} \omega_i.$$

Por consiguiente basta probar que

$$\int_{V_{p_i}} d\omega_i = \int_{V_{p_i}\cap\partial S} \omega_i,$$

es decir, hemos reducido el problema al caso local en que el soporte de la forma está contenido en el rango de una carta. Por simplificar la notación eliminaremos los subíndices, que son ya innecesarios. Hemos de probar que

$$\int_{V_p} d\omega = \int_{V_p\cap\partial S} \omega, \tag{10.3}$$

donde ω es una $n-1$-forma con soporte compacto contenido en V_p. Por linealidad podemos suponer además que

$$\omega = f\, dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_n.$$

Llamemos

$$\omega^* = (X_p \circ f)\, dx_1 \wedge \cdots \wedge dx_{i-1} \wedge dx_{i+1} \wedge \cdots \wedge dx_n,$$

que es una $n-1$-forma definida en U_p cuyo soporte es la antiimagen por el homeomorfismo X_p del soporte de ω, el cual estará contenido en la antiimagen de V_p, que es el interior del cubo C_p.

Como en teorema anterior, vemos que

$$d\omega = (-1)^{i-1} \frac{\partial f}{\partial x_i}\, dx_1 \wedge \cdots \wedge dx_n.$$

Para calcular el primer miembro de (10.3) transformamos la integral mediante la carta X_p, con lo que obtenemos

$$\int_{C_p} (-1)^{i-1} \left(X_p \circ \frac{\partial f}{\partial x_i} \right) dx_1 \wedge \cdots \wedge dx_n.$$

Teniendo en cuenta la definición de la derivada parcial de una función en una variedad es claro que

$$X_p \circ \frac{\partial f}{\partial x_i} = \frac{\partial (X_p \circ f)}{\partial x_i},$$

y al calcular igualmente $d\omega^*$ obtenemos

$$\int_{V_p} d\omega = \int_{C_p} d\omega^*.$$

Con esto podemos probar (10.3) en el caso en que p es un punto interior de S. En efecto, entonces V_p no corta a ∂S, luego el segundo miembro es nulo. Por otra parte, el soporte de ω^* está contenido en el interior del cubo C_p, luego ω^* es nula en ∂C_p, luego el teorema de Stokes para un cubo nos da que

$$\int_{V_p} d\omega = \int_{C_p} d\omega^* = \int_{\partial C_p} \omega^* = 0.$$

Así pues, en adelante supondremos que $p \in \partial S$. Para evaluar el segundo miembro de (10.3) usamos la carta positiva $X(0, x_2, \ldots, x_n)$. Notemos que en $V_p \cap \partial S$ la función x_1 es constante, luego $dx_1 = 0$. Supongamos primero $i = 1$. La carta transforma $V_p \cap \partial S$ en la cara $(C_p)_1^1$ del cubo, luego

$$\int_{V_p \cap \partial S} \omega = \int_{(C_p)_1^1} f\big(X_p(0, x_2, \ldots, x_n)\big)\, dx_2 \cdots dx_n = \int_{(C_p)_1^1} \omega^*,$$

(la última igualdad por definición de integral sobre una cara). La igualdad

$$\int_{V_p \cap \partial S} \omega = \int_{(C_p)_1^1} \omega^*$$

es válida aunque sea $i \neq 1$, pues en tal caso ω es nula en $V_p \cap \partial S$ y el miembro derecho es nulo por definición. En definitiva, la igualdad (10.3) que tenemos que probar se reduce a

$$\int_{C_p} d\omega^* = \int_{(C_p)_1^1} \omega^*.$$

Por el teorema de Stokes para un cubo basta probar que

$$\int_{\partial C_p} \omega^* = \int_{(C_p)_1^1} \omega^*.$$

Ahora bien, ω^* tiene el soporte contenido en la antiimagen por X_p de V_p, que es $]-1, 0] \times]-1, 1[^{n-1}$, lo que significa que ω^* es nula sobre todas las caras de C_p salvo quizá $(C_p)_1^1$, y aplicando la definición de integral sobre la frontera de un cubo tenemos la igualdad anterior. ■

El teorema anterior engloba a muchos casos particulares conocidos desde mucho antes, uno de ellos el Teorema de Stokes propiamente dicho, que se obtiene al aplicarlo al elemento de circulación de un campo en $\mathbb{R}^3$ a través de una curva.

Teorema 10.14 (Teorema de Stokes) *Sea S una superficie compacta orientable contenida en un abierto $U \subset \mathbb{R}^3$ y sea $F : U \longrightarrow \mathbb{R}^3$ un campo vectorial. Entonces*

$$\int_S (\operatorname{rot} F) n \, d\sigma = \int_{\partial S} F \, d\vec{r},$$

donde n es el vector normal a S que determina su orientación.

En otras palabras: el flujo del rotacional a través de la superficie es igual a su circulación en la frontera. Es consecuencia inmediata de la relación $d(F\,d\vec{r}) = d\Phi(\operatorname{rot} F)$, que demostramos en la sección anterior. La compacidad de la superficie implica la del soporte de la forma.

Consideremos ahora la relación $d(d\Phi(F)) = \operatorname{div} F\,dm$, que demostramos en la sección anterior. Al aplicarle el teorema de Stokes generalizado obtenemos otro importante teorema clásico debido a Gauss:

Teorema 10.15 (Teorema de la divergencia) *Sea $V \subset \mathbb{R}^m$ una variedad compacta de dimensión m contenida en un abierto U. Sea $F : U \longrightarrow \mathbb{R}^m$ un campo vectorial. Entonces*

$$\int_V \operatorname{div} F\,dm = \int_{\partial V} Fn\,d\sigma,$$

donde n es el vector normal a ∂V que apunta hacia fuera de V.

En otros términos, el flujo de un campo a través de una superficie cerrada es igual a la integral de la divergencia sobre el recinto que limita.

Ejemplo El campo $F(x) = x$ cumple $\operatorname{div} F = n$, luego el teorema de la divergencia nos da una fórmula para el volumen n-dimensional V encerrado por una superficie S:

$$V = \frac{1}{n}\int_S d\Phi(x).$$

Destacamos los casos particulares $n = 2$ y $n = 3$. El área de una figura plana limitada por una curva C es

$$A = \frac{1}{2}\int_C x\,dy - y\,dx.$$

El volumen de una región del espacio limitado por una superficie S es

$$V = \frac{1}{3}\int_S x\,dy \wedge dz + y\,dz \wedge dx + z\,dx \wedge dy.$$

Por ejemplo, la elipse de semiejes a y b admite la parametrización $x = a\cos t$, $y = b\operatorname{sen} t$. Por consiguiente su área es

$$A = \frac{1}{2}\int_0^{2\pi} (ab\cos^2 t + ab\operatorname{sen}^2 t)\,dt = \pi ab.$$

■

Ejercicio: Calcular el área de la cardioide mediante la fórmula anterior.

Ejemplo Consideremos el campo $F : \mathbb{R}^{n+1} \longrightarrow \mathbb{R}^{n+1}$ dado por

$$F(x) = \frac{x}{r},$$

para un $r > 0$. Sobre los puntos de la esfera de centro 0 y radio r coincide con el vector normal unitario de la misma, luego $Fn = 1$. Así pues, el teorema de la divergencia nos da que el área de la esfera vale

$$\int_S d\sigma = \int_S Fn\, d\sigma = \int_B \operatorname{div} F\, dm = \int_B \frac{n+1}{r}\, dm = (n+1)v_{n+1}r^n = \sigma_n r^n,$$

donde B es la bola de centro 0 y radio r y v_{n+1} es el volumen de la bola unitaria de dimensión $n+1$. Obtenemos de nuevo la relación $\sigma_n = (n+1)v_{n+1}$, que ya habíamos obtenido en el capítulo anterior. ∎

10.4 Aplicaciones del teorema de Stokes

El rotacional en hidrodinámica El teorema clásico de Stokes nos da una interpretación importante del rotacional en hidrodinámica. Supongamos que V es el campo de velocidades de un fluido tal y como considerábamos en el capítulo anterior. Supongamos que en su seno situamos una pequeña rueda que pueda girar en torno a un eje fijo. Formalmente, sea S un disco cerrado de centro p y radio r (contenido en el dominio de V, o sea, en $\mathbb{R}^3$, con cualquier inclinación). Sea C la circunferencia que lo bordea y sea n un vector unitario normal al mismo.

Dado $\epsilon > 0$, existe un entorno U de p tal que $|(\operatorname{rot} V)(p)n - (\operatorname{rot} V)(x)n| < \epsilon$ para todo $x \in U$. Si tomamos r suficientemente pequeño para que la rueda esté contenida en U, entonces

$$\big((\operatorname{rot} V)(p)n - \epsilon\big)\pi r^2 \leq \int_S (\operatorname{rot} F)n\, d\sigma \leq \big((\operatorname{rot} V)(p)n - \epsilon\big)\pi r^2.$$

Por otra parte, vimos en el capítulo anterior que

$$\int_C V\, d\vec{r} = 2\pi r^2 \omega_r,$$

donde ω_r es la velocidad angular que el fluido transmite a la rueda. Aplicando el teorema de Stokes llegamos a que

$$|(\operatorname{rot} V)(p)n - 2\omega_r| \leq \epsilon,$$

para todo r suficientemente pequeño, es decir,

$$(\operatorname{rot} V)(p)n = 2 \lim_{r \to 0} \omega_r.$$

Así pues, la velocidad angular que adquirirá la rueda es (aproximadamente) la mitad de la proyección del rotacional sobre el eje de giro. Claramente el rotacional indica la dirección en que hemos de situar el eje para que la velocidad de rotación sea máxima. ∎

La ecuación de continuidad Sea V el campo de velocidades de un fluido y sea ρ su densidad (ambos dependen de la posición y del tiempo). Sea p un punto cualquiera y S una esfera de centro p. En el capítulo anterior vimos que el flujo del campo $A = \rho V$ a través de S se interpreta como la masa de fluido que sale de S por unidad de tiempo. La cantidad de fluido contenida en S en un instante dado es la integral de ρ sobre la bola B de frontera S, luego la variación de esta masa es

$$\frac{d}{dt}\int_B \rho\, dm = \int_B \frac{\partial \rho}{\partial t}\, dm.$$

Sea r el radio de B y

$$\psi_r(p) = \frac{1}{m(B)}\left(\int_B \frac{\partial \rho}{\partial t}\, dm + \int_B \operatorname{div} A\, dm\right).$$

Así, $\psi_r(p)m(B)$ es el aumento de la masa de fluido en B por unidad de tiempo menos la cantidad de masa que entra en B a través de S por unidad de tiempo. Por consiguiente $\psi_r(p)$ es la cantidad de masa que se crea en B por unidad de tiempo y de volumen (la masa que aparece en B sin entrar por su frontera). El mismo argumento que hemos empleado en la interpretación del rotacional nos da ahora que

$$\psi(p) = \lim_{r\to 0} \psi_r(p) = \frac{\partial \rho}{\partial t}(p) + \operatorname{div} A(p),$$

donde $\psi(p)$ representa la cantidad de fluido que se crea alrededor de p por unidad de tiempo y de volumen. La ecuación

$$\operatorname{div} A = \psi - \frac{\partial \rho}{\partial t}. \tag{10.4}$$

se denomina *ecuación de continuidad de la hidrodinámica,* y expresa la conservación de la masa.

Los puntos donde $\psi > 0$ se llaman *fuentes* (son puntos donde aparece fluido) y los puntos donde $\psi < 0$ se llaman *sumideros* (en los cuales desaparece fluido).

Si ρ es constante el fluido se llama *incompresible* y entonces la ecuación fundamental se reduce a que $\rho \operatorname{div} V$ en un punto p es igual a la cantidad de fluido que se crea alrededor de p por unidad de masa y de volumen.

Si no hay fuentes ni sumideros la conservación de la masa se traduce en la ecuación $\operatorname{div} V = 0$. ■

La ecuación de Euler Veamos ahora la versión hidrodinámica de la segunda ley de Newton. Recordemos que ésta afirma que $F = ma$, donde F es la fuerza total que actúa sobre un cuerpo de masa m y a es la aceleración del mismo. Vamos a aplicarla a un elemento infinitesimal de fluido. Si llamamos $V(x,t)$ a la velocidad del fluido en el punto x y en el instante t, entonces una partícula de fluido sigue una trayectoria $X(t)$, de modo que $X'(t) = V(X(t),t)$. La aceleración $a(x,t)$ será

$$a(x,t) = X''(t) = \frac{\partial V}{\partial t} + \frac{\partial V}{\partial x}\frac{dx}{dt} + \frac{\partial V}{\partial y}\frac{dy}{dt} + \frac{\partial V}{\partial z}\frac{dz}{dt} = \frac{\partial V}{\partial t} + (\nabla V)V.$$

(Se entiende que ∇V se toma respecto a (x, y, z), pero no respecto a t).

La segunda ley de Newton para el elemento de fluido es $dF = \rho a\, dm$, donde dm es el elemento de volumen y ρ la densidad, de modo que $\rho\, dm$ es el elemento de masa. Si Ω es un volumen de fluido se ha de cumplir

$$F = \int_\Omega \rho \left(\frac{\partial V}{\partial t} + (\nabla V)V \right) dm,$$

donde F es la fuerza total que actúa sobre Ω. Podemos descomponer esta fuerza en dos partes. Por un lado tenemos la fuerza exterior que actúa sobre el fluido (por ejemplo la gravedad), cuya intensidad por unidad de masa representaremos por G, es decir, la fuerza exterior que actúa sobre Ω es $\int_\Omega \rho G\, dm$.

Además sobre Ω puede ejercerse una presión. Ésta puede deberse al fluido circundante, al recipiente (que impide que el fluido se derrame), a la presión atmosférica (si el recipiente está abierto por la parte superior), etc. Una característica de la presión es que se ejerce en todas direcciones. Más precisamente, la presión que el fluido ejerce sobre un cuerpo sumergido en el mismo (o sobre una porción del propio fluido) actúa en cada punto perpendicularmente a su superficie y hacia el interior. Si llamamos p (magnitud escalar) a la intensidad de la presión por unidad de superficie, entonces la presión total sobre V es $-\int_S pn\, d\sigma$, donde S es la superficie que limita a Ω y n es el vector normal a S que apunta hacia fuera de Ω. La componente i-ésima de esta integral es el flujo del campo pe_i. Por el teorema de la divergencia equivale a la integral en Ω de la derivada de p respecto a x_i, y al reunir las tres igualdades queda

$$-\int_S pn\, d\sigma = -\int_\Omega \nabla p\, dm.$$

Tenemos, pues

$$\int_\Omega \rho \left(\frac{\partial V}{\partial t} + (\nabla V)V \right) dm = \int_\Omega \rho G\, dm - \int_\Omega \nabla p\, dm.$$

Como esta igualdad ha de darse para todo volumen Ω, necesariamente los integrandos han de ser iguales, es decir,

$$\frac{\partial V}{\partial t} + (\nabla V)V = G - \frac{1}{\rho}\nabla p.$$

Ésta es la *ecuación de Euler,* que expresa la conservación de la cantidad de movimiento de un fluido.

Por ejemplo, si el fluido está en reposo, su densidad es constante y está sometido a un campo gravitatorio dirigido hacia abajo y de intensidad constante g, entonces la ecuación de Euler se reduce a

$$\nabla p = -\rho g e_3,$$

de donde se sigue fácilmente que $p = \rho g h$, donde h es la profundidad en el fluido (suponiendo que la presión en la superficie es nula).

Ejercicio: En las condiciones anteriores, Probar que un cuerpo sumergido en un fluido experimenta un empuje hacia arriba igual al peso del fluido que desplaza (principio de Arquímedes).

Veamos una aplicación de la ecuación de Euler. Sea C una curva cerrada contenida en el fluido. Sea $X(u)$ una parametrización. Sea $X(u,t)$ la posición en el instante t de la partícula de fluido que en un instante t_0 estaba en $X(u)$, es decir, $X(u,t_0) = X(u)$ y la derivada de X respecto de t es $V(X(u,t),t)$. Sea

$$\gamma(t) = \int_C V(x,t)\, d\vec{r} = \int_a^b V(X(u,t),t)\frac{\partial X}{\partial u}\, du.$$

Entonces

$$\frac{d\gamma}{dt} = \int_a^b \frac{dV}{dt}\frac{\partial X}{\partial u}\, du + \int_a^b V\frac{\partial}{\partial u}\frac{\partial X}{\partial t}\, du = \int_C \frac{dV}{dt}\, d\vec{r} + \frac{1}{2}\int_C d(VV).$$

Como C es una curva cerrada, la segunda integral es nula. En la primera aplicamos la ecuación de Euler:

$$\frac{d\gamma}{dt} = \int_C \left(G - \frac{1}{\rho}\nabla p\right) d\vec{r}.$$

Si suponemos que el campo de fuerzas exteriores es conservativo, entonces la integral de G a lo largo de C es nula y queda

$$\frac{d\gamma}{dt} = -\int_C \frac{1}{\rho}\nabla p\, d\vec{r} = \int_C \frac{dp}{\rho}.$$

Finalmente, si suponemos que la densidad en cada punto y en cada instante depende sólo de la presión, la última integral es nula (si $f(p)$ es una primitiva de $1/\rho(p)$ entonces el integrando es df). Con esto hemos probado:

> **Teorema de Lord Kelvin** *En un fluido sometido a un campo de fuerzas conservativo y en el que la densidad en cada punto e instante dependa sólo de la presión, la circulación de la velocidad a lo largo de cualquier curva cerrada permanece constante cuanto ésta se desplaza siguiendo la trayectoria del fluido.*

■

El gradiente de un campo escalar tiene una interpretación muy simple que no requiere el teorema de Stokes: dado un campo escalar ϕ en $\mathbb{R}^n$ y un punto p donde $\nabla\phi(p) \neq 0$, entonces para cada $v \in \mathbb{R}^n$ de norma 1 tenemos que

$$d\phi(p)(v) = \nabla\phi(p)v = \|\nabla\phi(p)\| \cos\alpha,$$

donde α es el ángulo entre v y $\nabla\phi(p)$. Esta cantidad es, por otra parte, el aumento que experimenta ϕ por cada unidad que nos desplazamos en la dirección de v y obviamente se hace máxima cuando v tiene la misma dirección y sentido que $\nabla\phi(p)$. Por consiguiente $\nabla\phi$ indica en cada punto la dirección en la que ϕ crece más rápidamente.

Definición 10.16 Sea $\phi : U \longrightarrow \mathbb{R}$ un campo escalar en un abierto de $\mathbb{R}^n$. Se llama *laplaciano* de ϕ a

$$\Delta\phi = \operatorname{div} \nabla\phi = \frac{\partial^2\phi}{\partial x_1^2} + \cdots + \frac{\partial^2\phi}{\partial x_n^2}.$$

Si S es una variedad $n-1$-dimensional orientable contenida en U y n es su vector normal, se define la *derivada direccional* de ϕ respecto a n como

$$\frac{d\phi}{dn} = (\nabla\phi)n.$$

(Notar que efectivamente se trata de la derivada direccional de ϕ en el sentido usual y en la dirección que marca n.) Aplicando a $\nabla\phi$ el teorema de la divergencia obtenemos:

Teorema 10.17 *Sea $V \subset \mathbb{R}^n$ una variedad compacta de dimensión n contenida en un abierto U. Sea $\phi : U \subset \mathbb{R}^n \longrightarrow \mathbb{R}$ un campo escalar. Entonces*

$$\int_V \Delta\phi \, dm = \int_{\partial V} \frac{d\phi}{dn} \, d\sigma,$$

donde n es el vector normal a ∂V que apunta hacia fuera de V.

En otras palabras, el flujo del gradiente de un campo a través de una superficie cerrada es igual a la integral de su laplaciano sobre el recinto que ésta encierra. Esto nos relaciona el laplaciano con los campos conservativos, que se expresan como gradientes de campos escalares.

El campo gravitatorio Consideremos de nuevo el campo gravitatorio generado por un cuerpo puntual de masa M situado en un punto y. Sabemos que su intensidad (es decir, la fuerza que ejerce por unidad de masa) viene dada por la ley de Newton:

$$E(x) = -\frac{GM}{\|x-y\|^3}\,(x-y),$$

suponiendo al cuerpo en el origen de coordenadas, y que además puede expresarse de la forma $E = -\nabla V$, donde

$$V(x) = -\frac{GM}{\|x-y\|}.$$

Un cálculo elemental muestra que $\Delta V = \operatorname{div} E = 0$. Esto significa que el flujo de E a través de una superficie cerrada S que rodee a un punto dado y no contenga a y es nulo (es la integral del laplaciano de V). No ocurre lo mismo si la superficie contiene a y en su interior (notar que si y está en S el flujo no está definido). En efecto, en tal caso podemos tomar una bola B de centro y contenida en la región G rodeada por S. Entonces $G \setminus B$ es una variedad con frontera, la cual es igual a $S \cup \partial B$. La orientación positiva en S es la misma respecto a G y respecto a $G \setminus B$, mientras que la orientación positiva en ∂B

como parte de la frontera de $G \setminus B$ es la dada por el vector normal que apunta hacia dentro de B, es decir, la opuesta a su orientación positiva como frontera de B. El teorema de Stokes nos da que el flujo de E por la frontera de $G \setminus B$ es nulo, luego el flujo a través de S es igual al flujo a través de ∂B. Por otra parte, el campo E es normal a ∂B, de módulo constante sobre la superficie y apunta hacia el interior, luego dicho flujo es

$$\Phi = -\|E\|m(\partial B) = -\frac{GM}{r^2}\,4\pi r^2 = -4\pi GM,$$

donde r es el radio de B.

Resulta orientador pensar en el campo gravitatorio generado por una masa puntual M como si fuera el campo de velocidades de un fluido incompresible. El hecho de que el flujo a través de las superficies cerradas que contienen a y sea $-4\pi GM$ se interpreta como que tales superficies "se tragan" una cantidad de fluido proporcional a M. No podemos decir con propiedad que y sea un sumidero en el sentido que dimos a este término, pues sobre él no está definido el campo, pero esto más bien debe llevarnos a pensar que dicha definición de sumidero es demasiado particular, y que debemos admitir como tales a los puntos alrededor de los cuales desaparece fluido en el sentido que acabamos de ver.

Por otra parte, donde no hay masa la divergencia del campo es nula y, efectivamente, no se crea ni se destruye fluido (el flujo es nulo).

Todo esto se generaliza de forma inmediata al caso del campo producido por n partículas puntuales de masas $M_1, \ldots, M_n$. Basta aplicar el principio de superposición, en virtud del cual la fuerza que estas masas ejercen sobre un cuerpo dado es la suma de las fuerzas que cada una de ellas ejercería por separado. Es claro entonces que el potencial del campo es la suma de los potenciales asociados a cada uno de ellos. El flujo a través de una superficie cerrada es igual a $-4\pi G$ por la suma de las masas que contiene.

Los modelos de masas puntuales son válidos para estudiar el movimiento de los planetas, pero no sirven para dar cuenta, por ejemplo, de la interacción entre la Tierra y los objetos próximos a ella. En tal caso debemos tener en cuenta la forma geométrica del espacio que ocupan las masas. En lugar de tener uno o varios puntos de masa tenemos una medida que asigna a cada región del espacio la masa que contiene. Si admitimos que una región de volumen 0 no puede contener masa (es decir, negamos la existencia de masas puntuales) entonces dicha medida estará determinada por una función de densidad ρ, de modo que la masa contenida en un volumen V vendrá dada por

$$M = \int_V \rho\, dm.$$

Para calcular una aproximación de la intensidad del campo gravitatorio generado por la distribución de masas ρ en un punto x podemos dividir el espacio en regiones pequeñas de volumen $\rho\, dm$, calcular la intensidad correspondiente a esta masa y sumar las fuerzas así obtenidas. Con ello estamos calculando una

aproximación de la integral

$$E(x) = -G \int_V \frac{\rho(y)}{\|x-y\|^3}\,(x-y)\,dm(y),$$

donde V es un volumen que contiene a toda la masa que influye (el dominio de ρ) y la integral de una función vectorial se interpreta como el vector formado por la integral de cada componente. Ésta debe ser la expresión exacta de la citada intensidad del campo. Un razonamiento similar con los potenciales nos lleva a que el potencial gravitatorio en el punto x debe ser

$$V(x) = -G \int_V \frac{\rho(y)}{\|x-y\|}\,dm(y).$$

No obstante, todo esto nos plantea varios problemas. En primer lugar hemos de justificar que las integrales existen, pues si $x \in V$ el integrando tiende a infinito en x. Por otra parte no es evidente que estas funciones cumplan $E = -\nabla V$, que es la relación que debe darse para que V sea una función potencial de E. Los resultados que vamos a obtener pueden darse en un contexto general:

Definición 10.18 Sea $\Omega \subset \mathbb{R}^n$ (con $n \geq 3$) un abierto acotado y f una función medible acotada en Ω. Llamaremos *potencial newtoniano* asociado a f a la función

$$V_f(x) = \int_\Omega \frac{f(y)}{\|x-y\|^{n-2}}\,dm(y), \quad \text{para } x \in \mathbb{R}^n.$$

El teorema 9.23 garantiza que el integrando es realmente integrable en Ω, por lo que V_f está bien definido. Sea $g(x,y) = \|x-y\|^{2-n}$. Es claro que g es de clase C^1 en $(\mathbb{R}^n \setminus \overline{\Omega}) \times \overline{\Omega}$, luego el teorema 7.23 nos garantiza[2] que V_f es de clase C^1 en $\mathbb{R}^n \setminus \overline{\Omega}$ y sus derivadas valen

$$\frac{\partial V_f}{\partial x_i}(x) = -(n-2)\int_\Omega \frac{f(y)}{\|x-y\|^n}\,(x_i-y_i)\,dm(y), \tag{10.5}$$

Vamos a probar que esta expresión vale igualmente en $\overline{\Omega}$. Por lo pronto observemos que el integrando del segundo miembro es ciertamente integrable. Basta tener en cuenta que

$$\frac{|x_i-y_i|}{\|x-y\|} \leq 1,$$

con lo que el integrando está mayorado por la función integrable $K/\|x-y\|^{n-1}$.

Para cada natural $k \geq 1$ consideremos la función $a_k : [0,+\infty[\longrightarrow \mathbb{R}$ dada por

$$a_k(r) = \begin{cases} \dfrac{1}{k^{n-2}} + \dfrac{n-2}{k^{n-3}}\left(r - \dfrac{1}{k}\right) & \text{si } r < 1/k \\ r^{n-2} & \text{si } r \geq 1/k \end{cases}$$

[2]Admitimos que $\partial\Omega$ es nula, luego la integral puede tomarse en el compacto $\overline{\Omega}$.

Claramente a_k es de clase C^1 en su dominio y además no se anula, pues la derivada es positiva en $[0, 1/k]$. Definimos las funciones

$$V_k(x) = \int_\Omega \frac{f(y)}{a_k(\|x-y\|)}\, dm(y).$$

La función $a_k(\|x-y\|)^{-1}$ es de clase C^1 en $\mathbb{R}^n \times \mathbb{R}^n$, por lo que podemos aplicar el teorema 7.23.

Si probamos que las funciones V_k convergen uniformemente a V_f y sus derivadas convergen al segundo miembro de (10.5), el teorema 3.28 nos dará que dicha igualdad es válida en todo punto.

Puesto que los integrandos de $V_f(x)$ y $V_k(x)$ difieren sólo sobre $B_{1/k}(x)$, tenemos que

$$\begin{aligned} |V_f(x) - V_k(x)| &\leq M \int_{B_{1/k}(x)} \left| \frac{1}{\|x-y\|^{n-2}} - \frac{1}{a_k(\|x-y\|)} \right| dm(y) \\ &= M \int_{B_{1/k}(0)} \left| \frac{1}{\|y\|^{n-2}} - \frac{1}{a_k(\|y\|)} \right| dm(y). \end{aligned}$$

La última integral, como función del dominio de integración, es una medida finita en $B_1(0)$ por el teorema 7.17, luego el último miembro tiende a 0 con k, por el teorema 7.2. Esto prueba la convergencia uniforme de V_k. El mismo argumento vale para las derivadas. Observar que no es necesario calcular explícitamente la derivada del integrando de V_k. Basta tener en cuenta que consta de $f(y)$ multiplicada por una función continua, luego integrable.

Notemos que las derivadas de V_f en los puntos de Ω son el límite uniforme de una sucesión de funciones continuas (las derivadas de V_k), luego V_f es una función de clase C^1 en $\mathbb{R}^n$.

En particular tenemos que el campo y el potencial gravitatorio determinados por una distribución de masa ρ están bien definidos y satisfacen la relación $E = -\nabla V$, como ha de ser.

Sigamos en el caso general y vamos a calcular el laplaciano de V_f. El teorema 7.23 nos permite concluir directamente que V_f es de clase C^∞ en $\mathbb{R}^n \setminus \overline{\Omega}$. Más aún, es fácil comprobar que $\Delta_x g = 0$, con lo que también $\Delta V_f = 0$. Para los puntos de $\overline{\Omega}$ no podemos emplear la misma técnica que hemos usado para calcular las primeras parciales, pues las derivadas segundas del integrando no son integrables.

Consideremos un punto $x_0 \in \Omega$ tal que f es de clase C^1 en una bola $B_{2\epsilon}(x_0) \subset \Omega$.

Descomponemos $V_f = V_1 + V_2$, donde ambos sumandos tienen la misma definición que V_f salvo que el dominio de integración es $\overline{B}_\epsilon(x_0)$ en el caso de V_1 y $\overline{\Omega} \setminus B_\epsilon(x_0)$ en el caso de V_2.

Es claro que V_2 es de clase C^2 en $B_\epsilon(x_0)$. Sus parciales segundas se pueden calcular derivando el integrando. Además $\Delta V_2 = 0$. Por lo tanto para probar

que V_f es de clase C^2 en $B_\epsilon(x_0)$ basta probar que lo es V_1, y además tendremos $\Delta V_f(x_0) = \Delta V_1(x_0)$. Ya sabemos que

$$\begin{aligned}
\frac{\partial V_1}{\partial x_i}(x) &= \int_{\overline{B}_\epsilon(x_0)} f(y)\,\frac{\partial}{\partial x_i}\left(\frac{1}{\|x-y\|^{n-2}}\right) dm(y) \\
&= -\int_{\overline{B}_\epsilon(x_0)} f(y)\,\frac{\partial}{\partial y_i}\left(\frac{1}{\|x-y\|^{n-2}}\right) dm(y) \\
&= -\int_{\overline{B}_\epsilon(x_0)} \left(\frac{\partial}{\partial y_i}\left(\frac{f(y)}{\|x-y\|^{n-2}}\right) - \frac{1}{\|x-y\|^{n-2}}\frac{\partial f}{\partial y_i}(y)\right) dm(y).
\end{aligned}$$

La integral del segundo término es el potencial newtoniano de la derivada de f, luego sabemos que tiene derivada continua y viene dada por

$$\frac{\partial}{\partial x_i}\int_{\overline{B}_\epsilon(x_0)} \frac{1}{\|x-y\|^{n-2}}\frac{\partial f}{\partial y_i}(y)\,dm(y) = \int_{\overline{B}_\epsilon(x_0)} \frac{\partial}{\partial x_i}\left(\frac{1}{\|x-y\|^{n-2}}\right)\frac{\partial f}{\partial y_i}(y)\,dm(y)$$

Nos ocupamos ahora del otro término. Aplicamos el teorema de la divergencia al campo F dado por

$$F(y) = \frac{f(y)}{\|x-y\|^{n-2}}\,e_i,$$

donde e_i es el i-ésimo vector de la base canónica. La divergencia de F es nuestro integrando y su flujo a través de la esfera de radio ϵ (precedido del signo negativo de nuestra integral) es

$$-\int_{\partial B_\epsilon(x_0)} \frac{f(y)}{\|x-y\|^{n-2}}\,e_i n(y)\,d\sigma(y),$$

donde n es el vector unitario normal a la esfera y $d\sigma$ es el elemento de medida de la esfera. A esta integral también le podemos aplicar 7.23, con lo que tiene derivada continua respecto a x_i y viene dada por

$$(n-2)\int_{\partial B_\epsilon(x_0)} \frac{f(y)}{\|x-y\|^{n}}\,(x_i-y_i)e_i n(y)\,d\sigma(y).$$

En este punto ya tenemos que V_1 es de clase C^2 en $B_\epsilon(x_0)$. Teniendo en cuenta que $n(y) = (y-x_0)/\|x_0-y\|$, al particularizar en x_0 tenemos en total

$$\begin{aligned}
\frac{\partial^2 V_1}{\partial x_i^2}(x_0) &= -(n-2)\int_{\partial B_\epsilon(x_0)} \frac{f(y)(x_{0i}-y_i)^2}{\epsilon^{n+1}}\,d\sigma(y) \\
&+ \int_{\overline{B}_\epsilon(x_0)} \frac{\partial}{\partial x_i}\left(\frac{1}{\|x-y\|^{n-2}}\right)(x_0,y)\frac{\partial f}{\partial y_i}(y)\,dm(y).
\end{aligned}$$

Por consiguiente:

$$\begin{aligned}
\Delta V_f(x_0) = \Delta V_1(x_0) &= -\frac{n-2}{\epsilon^{n-1}}\int_{\partial B_\epsilon(x_0)} f(y)\,d\sigma \\
&+ \sum_{i=1}^n \int_{\overline{B}_\epsilon(x_0)} \frac{\partial}{\partial x_i}\left(\frac{1}{\|x-y\|^{n-2}}\right)(x_0,y)\frac{\partial f}{\partial y_i}\,dm.
\end{aligned}$$

Como el miembro izquierdo no depende de ϵ, podemos tomar el límite cuando ϵ tiende a 0. El último sumatorio tiende a 0, luego queda

$$\Delta V_f(x_0) = -(n-2) \lim_{\epsilon \to 0} \frac{1}{\epsilon^{n-1}} \int_{\partial B_\epsilon(x_0)} f(y)\, d\sigma.$$

Llamemos σ_{n-1} a la medida de $\partial B_1(0)$. Entonces la medida de $\partial B_\epsilon(x_0)$ es $\epsilon^{n-1}\sigma_{n-1}$. Así

$$\left| \frac{1}{\epsilon^{n-1}} \int_{\partial B_\epsilon(x_0)} f(y)\, d\sigma - \sigma_{n-1} f(x_0) \right| = \left| \frac{1}{\epsilon^{n-1}} \int_{\partial B_\epsilon(x_0)} f(y) - f(x_0)\, d\sigma \right|$$

$$\leq \frac{1}{\epsilon^{n-1}} \int_{\partial B_\epsilon(x_0)} |f(y) - f(x_0)|\, d\sigma.$$

Dado $\eta > 0$, existe un $\delta > 0$ de manera que si $\|y - x_0\| < \delta$ entonces $|f(y) - f(x_0)| < \eta/\sigma_{n-1}$. Para todo $\epsilon < \delta$ la expresión anterior está acotada por η, luego concluimos que $\Delta V_f(x_0) = -(n-2)\sigma_{n-1} f(x_0)$.

Si extendemos f a $\mathbb{R}^n$ con el valor 0 fuera de Ω, hemos probado que V_f es de clase C^2 allí donde f es de clase C^2 (lo cual incluye a todos los puntos de $\mathbb{R}^n \setminus \overline{\Omega}$) y en tales puntos $\Delta V_f = -(n-2)\sigma_{n-1} f$. El análisis que hemos hecho de las parciales de V_f puede usarse inductivamente para probar que si f es de clase C^k alrededor de un punto, lo mismo vale para V_f. Resumimos en un teorema lo que hemos obtenido:

Teorema 10.19 *Sea Ω un abierto acotado en $\mathbb{R}^n$, para $n \geq 3$, y sea f una función medible acotada que se anula fuera de Ω. Entonces el potencial newtoniano*

$$V_f(x) = \int_\Omega \frac{f(y)}{\|x - y\|^{n-2}}\, dm(y), \quad \text{para } x \in \mathbb{R}^n$$

es una función de clase C^1 en $\mathbb{R}^n$ y de clase C^k en todos los puntos donde f es de clase C^k. Además

$$\nabla V_f = -(n-2) \int_\Omega \frac{f(y)}{\|x - y\|^n} (x - y)\, dm(y),$$

y en los puntos donde f es de clase C^2 satisface la ecuación de Poisson *asociada a f, es decir, $\Delta V_f = -(n-2)\sigma_{n-1} f$, donde σ_{n-1} es la medida de la esfera unitaria de dimensión $n-1$.*

Ejercicio: Probar un teorema análogo para $n = 2$ definiendo

$$V_f(x) = \int_\Omega f(y) \log \|x - y\|\, dy.$$

Si $n = 3$ queda $\Delta V_f = -4\pi f$. En particular, si V es el potencial gravitatorio generado por una distribución de masa ρ, entonces $\Delta V = 4\pi G\rho$. Equivalentemente, si E es la intensidad del campo, se cumple $\operatorname{div} E = -4\pi G\rho$. En este caso

podemos decir con propiedad que los puntos donde hay masa se comportan como sumideros de un “fluido gravitatorio”. Además tenemos un importante teorema de Gauss:

> *El flujo del campo gravitatorio a través de una superficie cerrada que encierra una masa M es igual a $-4\pi GM$.*

Por ejemplo, sea B una esfera homogénea de densidad ρ y radio R. Es fácil ver que el campo gravitatorio que origina tiene simetría esférica, es decir,

$$E(x) = -\frac{K(\|x\|)}{\|x\|}\,x,$$

para una cierta función K. Consideremos una superficie esférica S cuyo centro coincida con el de B y de radio $r < R$. El flujo de E a través de S es $-4\pi r^2 K(r)$, luego por el teorema de Gauss

$$-4\pi r^2 K(r) = -4\pi G\rho\,\frac{4}{3}\pi r^3.$$

En definitiva,

$$E(x) = -\frac{4\pi G\rho}{3}\,x = -\frac{GM}{R^3}\,x.$$

Si tomamos $r \geq R$ entonces queda $-4\pi r^2 K(r) = -4\pi GM$, luego

$$E(x) = -\frac{GM}{\|x\|^3}\,x,$$

es decir, el campo generado por la esfera en un punto exterior a la misma coincide con el que generaría una masa puntual situada en su centro. ■

La ecuación del calor La materia, en cualquiera de sus estados, está compuesta de partículas diminutas, sean partículas subatómicas sueltas, átomos con enlace metálico o moléculas con diferentes estructuras. En todos estos casos, dichas partículas tienen una cierta libertad de movimiento y a nivel microscópico pueden moverse a velocidad considerable. Esta velocidad no puede medirse directamente, pero la velocidad media de las partículas de un cuerpo determina lo que llamamos su temperatura T. Por otra parte, la suma de la energía cinética de cada partícula es lo que llamamos la cantidad de calor Q del cuerpo (y se mide en Julios, como corresponde a la energía). Puesto que T es un promedio de velocidades, ya no es cierto que Q sea proporcional al cuadrado de T, sino que la experiencia establece que la proporción es lineal y la constante depende de las características químicas de cada sustancia. Concretamente cada sustancia tiene asociado un *calor específico* c, de modo que la cantidad de calor de un cuerpo de masa m, calor específico c y temperatura T es $Q = mcT$.

Aquí suponemos que T y c son constantes. Si c y T dependen de la posición entonces $Q = \int_V c\rho T\,dm$, donde ρ es la densidad del cuerpo (función de la posición) y dm es el elemento de volumen (no de masa). Por consiguiente $c\rho T$ es

la densidad de calor de un cuerpo de calor específico c, densidad ρ y temperatura T. Podemos suponer que c y ρ sólo dependen de la posición, mientras que Q y T dependerán también del tiempo, y se plantea el problema de determinar esta dependencia, esto es, de determinar la forma en que se transmite el calor a través de un cuerpo.

El modelo más simple al respecto postula que el calor es como un fluido que se mueve hacia el punto más frío posible, es decir, teniendo en cuenta el ejemplo en el que hemos introducido la ecuación de continuidad así como la interpretación del gradiente: $A = -k\nabla T$, donde $k > 0$ es una constante. La ecuación de continuidad (10.4) se convierte en este caso en

$$k\Delta T + \psi = c\rho\frac{\partial T}{\partial t},$$

donde ψ refleja las fuentes y sumideros de calor. Esta ecuación se conoce como *ecuación del calor.* ■

10.5 Las fórmulas de Green

Vamos a deducir varias fórmulas clásicas a partir del teorema de Stokes. Partimos de dos funciones $f, g : U \longrightarrow \mathbb{R}$ de clase C^2 en un abierto $U \subset \mathbb{R}^n$. Una simple comprobación nos da la identidad

$$\operatorname{div}(g\nabla f) = \nabla f\,\nabla g + g\Delta f.$$

Sea $V \subset U$ una variedad compacta orientable de dimensión n contenida en U. Aplicando el teorema de la divergencia obtenemos la llamada *primera fórmula de Green:*

$$\int_V g\Delta f\,dm + \int_V \nabla g\,\nabla f\,dm = \int_{\partial V} g\,\frac{df}{dn}\,d\sigma,$$

donde $d\sigma$ es el elemento de medida en ∂V y n su vector normal. Intercambiando los papeles de f y g y restando las fórmulas correspondientes obtenemos la *segunda fórmula de Green:*

$$\int_V (g\Delta f - f\Delta g)\,dm = \int_{\partial V}\left(g\,\frac{df}{dn} - f\,\frac{dg}{dn}\right)d\sigma.$$

Supongamos $n \geq 3$, fijemos un punto x interior a V y apliquemos la fórmula anterior a la función $g(y) = 1/\|x-y\|^{n-2}$ y a la variedad V_ϵ que resulta de quitarle a V una bola abierta de radio ϵ suficientemente pequeño para que esté contenida en V. Observamos que $\Delta g = 0$. En efecto:

$$\frac{\partial g}{\partial y_i} = (n-2)\frac{x_i - y_i}{\|x-y\|^n}, \quad \frac{\partial g^2}{\partial y_i^2} = -\frac{n-1}{\|x-y\|^n} + n(n-2)\frac{(x_i-y_i)^2}{\|x-y\|^{n+2}},$$

y al sumar sobre i queda 0. Si llamamos S_ϵ a la esfera de centro x y radio ϵ la fórmula de Green nos da que

$$\begin{aligned}\int_{V_\epsilon} \frac{\Delta f(y)}{\|x-y\|^{n-2}}\, dm(y) &= \int_{\partial V} \left(\frac{1}{\|x-y\|^{n-2}} \frac{df}{dn} - f\, \frac{d}{dn} \frac{1}{\|x-y\|^{n-2}} \right) d\sigma(y) \\ &+ \int_{S_\epsilon} \left(f\, \frac{d}{dn} \frac{1}{\|x-y\|^{n-2}} - \frac{1}{\|x-y\|^{n-2}} \frac{df}{dn} \right) d\sigma(y).\end{aligned}$$

Observar que hemos cambiado el signo en el segundo integrando porque la orientación positiva de S es la contraria a la que tiene como parte de la frontera de V_ϵ. Puesto que f es de clase C^2, tenemos que Δf es continua en V, por lo que el integrando del primer miembro es integrable en V (es el potencial newtoniano de Δf), luego existe el límite cuando $\epsilon \to 0$ de ambos miembros de la igualdad, luego también del último término. Vamos a calcularlo.

Notemos que sobre los puntos de S_ϵ es $n(y) = (y-x)/\|x-y\|$, luego los cálculos que hemos hecho antes muestran que

$$\frac{d}{dn} \frac{1}{\|x-y\|^{n-2}} = -\frac{n-2}{\|x-y\|^{n-1}}\,.$$

El último término es, pues,

$$-\frac{(n-2)}{\epsilon^{n-1}} \int_{S_\epsilon} f(y)\, d\sigma(y) - \int_{S_\epsilon} \frac{1}{\epsilon^{n-2}} \frac{df}{dn}\, d\sigma(y).$$

Si llamamos σ_{n-1} a la medida de Lebesgue de la esfera unitaria de dimensión $n-1$, entonces $\sigma(S_\epsilon) = \epsilon^{n-1}\sigma_{n-1}$, y es claro que la segunda integral está acotada por $K\sigma_{n-1}\epsilon$, donde K es una cota de la derivada direccional de f en un entorno de x. Por consiguiente este término tiende a 0. El límite del primer sumando lo calculamos al estudiar los potenciales newtonianos (ver las fórmulas precedentes al teorema 10.19). Recordemos que vale $-(n-2)\sigma_{n-1} f(x)$. Con esto obtenemos la *tercera fórmula de Green*

$$\begin{aligned}f(x) &= \frac{1}{(n-2)\sigma_{n-1}} \left(-\int_V \frac{\Delta f(y)}{\|x-y\|^{n-2}}\, dm(y) \right. \\ &+ \left. \int_{\partial V} \left(\frac{1}{\|x-y\|^{n-2}} \frac{df}{dn} - f\, \frac{d}{dn} \frac{1}{\|x-y\|^{n-2}} \right) d\sigma(y) \right).\end{aligned}$$

Esta fórmula nos dice que el valor de una función de clase C^2 en un punto x está completamente determinado por Δf en un entorno de x y las funciones f y df/dn sobre una superficie que rodee a x. Hay un caso particularmente interesante donde esta expresión se simplifica mucho:

Definición 10.20 Se dice que una función f de clase C^2 es *harmónica* en un abierto V si cumple $\Delta f(x) = 0$ para todo $x \in V$.

Por ejemplo, el potencial newtoniano de una función f es una función harmónica en los puntos exteriores al soporte de f.

La tercera fórmula de Green para una función harmónica f se reduce a

$$f(x) = \frac{1}{(n-2)\sigma_{n-1}} \int_{\partial V} \left(\frac{1}{\|x-y\|^{n-2}} \frac{df}{dn} - f \frac{d}{dn} \frac{1}{\|x-y\|^{n-2}} \right) d\sigma(y),$$

donde V es un abierto en $\mathbb{R}^n$ cuya clausura sea una variedad compacta y en el cual f sea harmónica (con $n \geq 3$).

Esta fórmula nos dice en principio que una función harmónica f en V está completamente determinada por los valores que f y df/dn toman sobre ∂V. Podemos ir más lejos y concluir que una función harmónica en V está completamente determinada por los valores que toma en ∂V. En efecto, supongamos que f_1 y f_2 son funciones continuas en $\overline{V}$, harmónicas en V y que coinciden en ∂V. Entonces la función $h = f_1 - f_2$ es harmónica en V y se anula en ∂V. Aplicando la primera fórmula de Green a las funciones $f = g = h$ resulta

$$\int_V \|\nabla h\|^2 \, dm = 0$$

y, como el integrando es positivo, ha de ser $\nabla h = 0$ en V, lo que implica que h es constante en V y, como se anula en ∂V, ha de ser $h = 0$, es decir, $f_1 = f_2$.

Si suponemos ahora que V es la bola de centro x y radio r, entonces la tercera fórmula de Green para una función harmónica nos da que

$$f(x) = \frac{1}{r^{n-2}(n-2)\sigma_{n-1}} \int_{\partial V} \frac{df}{dn} \, d\sigma + \frac{1}{r^{n-1}\sigma_{n-1}} \int_{\partial V} f \, d\sigma,$$

Por el teorema de la divergencia

$$\int_{\partial V} \frac{df}{dn} \, d\sigma = \int_{\partial V} \nabla f \, n \, d\sigma = \int_V \Delta f \, dm = 0,$$

luego se cumple el llamado *teorema del valor medio de Gauss:*

$$f(x) = \frac{1}{\sigma(\partial B_r(x))} \int_{\partial B_r(x)} f \, d\sigma.$$

Esta fórmula afirma que el valor que toma una función harmónica en un punto x es la media aritmética de los valores que toma en cualquier esfera con centro en x. De aquí se sigue fácilmente que una función harmónica no puede tomar valores máximos o mínimos en ningún abierto en el que esté definida. También es claro que si una función harmónica tiende a una constante en ∞ entonces es constante.

Ejemplo Si $f, g : \mathbb{R}^n \longrightarrow \mathbb{R}$, usaremos la notación $f = O(g)$ para indicar que existen constantes M y R tales que si $\|x\| \geq R$ entonces $|f(x)| \leq M|g(x)|$ (se dice entonces que f es una función del orden de g).

Si $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ es una función de clase C^2 con soporte compacto, donde $n \geq 3$, es fácil ver que su potencial newtoniano V_f cumple $V_f = O(1/\|x\|^{n-2})$

y $\|\nabla V_f\| = O(1/\|x\|^{n-1})$. Por ejemplo, en el caso de la gravedad esto significa que el campo gravitatorio se atenúa en proporción inversa al cuadrado de la distancia. Además sabemos que satisface la ecuación $\Delta V_f = -(n-2)\sigma_{n-1}f$. Vamos a probar que V_f es la única función que cumple estas condiciones.

Supongamos que $u : \mathbb{R}^n \longrightarrow \mathbb{R}$ es una función de clase C^2 tal que

$$u = O(1/\|x\|), \quad \|\nabla u\| = O(1/\|x\|^2), \quad \Delta u = -(n-2)\sigma_{n-1}f.$$

Veamos que necesariamente $u = V_f$. Basta aplicar la tercera fórmula de Green a la bola B_r de centro 0 y radio r:

$$\begin{aligned} u(x) &= \int_{B_r} \frac{f(y)}{\|x-y\|^{n-2}}\, dm(y) \\ &+ \frac{1}{(n-2)\sigma_{n-1}} \int_{\partial B_r} \left(\frac{1}{\|x-y\|^{n-2}} \frac{du}{dn} - u \frac{d}{dn} \frac{1}{\|x-y\|^{n-2}} \right) d\sigma(y) \end{aligned}$$

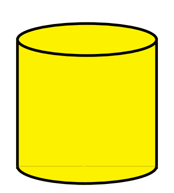
Si r es suficientemente grande como para que se cumplan las estimaciones de u y ∇u que estamos suponiendo, el módulo del integrando del último término está acotado por K/r^n, luego la integral está acotada por K'/r, luego tiende a 0 cuando r tiende a $+\infty$. Consecuentemente

$$u(x) = \int_{\mathbb{R}^n} \frac{f(y)}{\|x-y\|^{n-2}}\, dm(y).$$

El dominio de integración se puede reducir a cualquier abierto que contenga al soporte de f. Así pues, $u = V_f$. ■

Ejercicio: Deducir la tercera fórmula de Green y sus consecuencias para el caso en que $n = 2$, tomando para ello $g(y) = \log \|x-y\|$ en lugar de $\|x-y\|^{2-n}$.

10.6 El teorema de Stokes con singularidades



Observemos que el teorema de Stokes para un cubo (teorema 10.12) no es un caso particular del teorema de Stokes generalizado, pues un cubo no se ajusta a la definición que hemos dado de variedad con frontera (a causa de sus aristas). El hecho de que el teorema de Stokes valga para cubos hace sospechar que vale para variedades (en algún sentido de la palabra) más generales que las que estamos considerando aquí. Efectivamente, es frecuente que en aplicaciones a la física se haga uso del teorema sobre —por ejemplo— un cilindro de altura finita, que tampoco es una variedad con frontera a causa de las dos circunferencias que bordean sus "tapas". Podríamos considerar como variedad con frontera al cilindro menos dichas circunferencias, pero esto no ayuda en mucho, pues si tenemos una 2-forma definida en un entorno del cilindro su restricción al cilindro menos las circunferencias no tiene necesariamente soporte compacto (y nos gustaría, pese a ello, justificar la fórmula de Stokes en este caso). Conviene introducir algunos conceptos.

Definición 10.21 Sea $S \subset \mathbb{R}^m$ una variedad de dimensión n sin frontera. Sea $F(S) = \overline{S} \setminus S$. Diremos que un punto $p \in F(S)$ es un *punto frontera regular* de S si existe una carta $X : U \longrightarrow V$ de $\mathbb{R}^n$ alrededor de p (de coordenadas $x_1, \ldots, x_m$) de modo que $S \cap V$ está formado por los puntos de coordenadas $x_{n+1} = \cdots = x_m = 0$, $x_n < 0$, mientras que los puntos de $F(S) \cap V$ son los de coordenadas $x_n = x_{n+1} = \cdots = x_m = 0$. Llamaremos ∂S al conjunto de puntos frontera regulares de S.

Obviamente $S \cup \partial S$ es una variedad con frontera. El conjunto $F(S) \setminus \partial S$ es cerrado en $\mathbb{R}^m$. Sus puntos se llaman *puntos frontera singulares.*

Por ejemplo, si S es un cilindro abierto en $\mathbb{R}^3$, sus puntos frontera singulares son los de las dos circunferencias que limitan sus tapas. Nuestra intención es probar el teorema de Stokes para una variedad S cuyos puntos frontera singulares formen un conjunto pequeño en el sentido de la teoría de la medida. A su vez, la idea es modificar cada forma en un entorno suficientemente pequeño del conjunto de puntos singulares para hacer aplicable el teorema de Stokes que conocemos y después hacer un paso al límite.

Una *sucesión fundamental* de entornos de un cerrado $E \subset \mathbb{R}^m$ es una familia de abiertos $\{W_k\}_{k=1}^{\infty}$ que contienen a E tal que si V es un abierto y $E \subset V$, entonces $W_k \subset V$ para todo k suficientemente grande.

Supongamos que E es el conjunto de puntos singulares de una variedad S y que $\{W_k\}_{k=1}^{\infty}$ es una sucesión fundamental de entornos de E. Para cada k, tomemos una función g_k que se anule en un entorno de E y valga 1 fuera de W_k. De este modo, si ω es una $n-1$-forma definida en un entorno de S, la forma $g_k\omega$ coincide con ω salvo en W_k y tiene soporte compacto en $S \cup \partial S$, luego podemos aplicarle el teorema de Stokes:

$$\int_{\partial S} g_k\omega = \int_S d(g_k\omega) = \int_S g_k\, d\omega + \int_S dg_k \wedge \omega. \tag{10.6}$$

El paso siguiente es tomar límites cuando k tiende a infinito, y el punto más delicado es estudiar el comportamiento del último término. Recogeremos en una definición todo lo que necesitamos:

Definición 10.22 Sean E y S subconjuntos cerrados de $\mathbb{R}^m$. Diremos que E es *despreciable* para S si existe un abierto W en $\mathbb{R}^n$ que contiene a E y una sucesión fundamental $\{W_k\}_{k=1}^{\infty}$ de entornos de E tales que $\overline{W}_k \subset W$ y una sucesión $\{g_k\}_{k=1}^{\infty}$ de funciones de clase C^1 en W tales que

a) $0 \leq g_k \leq 1$, g_k se anula en un entorno de E y vale 1 fuera de W_k.

b) Si ω es una $n-1$-forma de clase C^1 en W, entonces $dg_k \wedge \omega$ es integrable en $W \cap S$ y, si llamamos μ_k a la medida signada en S definida por su integral, entonces
$$\lim_k |\mu_k|(W \cap S) = 0.$$

Con esta definición es fácil probar:

Teorema 10.23 *Sea $S \subset \mathbb{R}^m$ una variedad de dimensión n sin frontera. Sea ω una $n-1$-forma de clase C^1 en un abierto de $\mathbb{R}^m$ que contenga a $\overline{S}$ y tal que la intersección con $\overline{S}$ del soporte de ω sea compacta. Supongamos:*

a) Si E es la intersección del conjunto de puntos frontera singulares de S con el soporte de ω, entonces E es despreciable para S.

b) Las formas $d\omega$ en S y ω en ∂S son integrables.

Entonces

$$\int_S d\omega = \int_{\partial S} \omega.$$

Demostración: Sean W, $\{W_k\}_{k=1}^\infty$ y $\{g_k\}_{k=1}^\infty$ según la definición de conjunto despreciable. Notar que las funciones g_k se pueden considerar definidas en $\mathbb{R}^m$. Entonces $g_k\omega$ es nula en un entorno de E, de donde se sigue fácilmente que el soporte de su restricción a $S \cup \partial S$ es compacto. Aplicando el teorema de Stokes a esta variedad con frontera obtenemos (10.6). Ahora notamos que

$$\left| \int_{\partial S} \omega - \int_{\partial S} g_k\omega \right| = \left| \int_{\partial S} (1-g_k)\omega \right| \leq \int_{W_k \cap \partial S} d|\mu_\omega| = |\mu_\omega|(W_k \cap \partial S),$$

donde μ_ω es la medida definida por la integral de ω. Puesto que la intersección de los conjuntos $W_k \cap \partial S$ es vacía y las medidas son finitas, el teorema 7.2 nos da que

$$\lim_k \int_{\partial S} g_k\omega = \int_{\partial S} \omega.$$

(Podemos suponer que los conjuntos W_k son decrecientes.) Igualmente se llega a que

$$\lim_k \int_S g_k \, d\omega = \int_S d\omega.$$

Finalmente:

$$\left| \int_S dg_k \wedge \omega \right| \leq \int_{S \cap W} d|\mu_k| = |\mu_k|(W \cap S),$$

y por la definición de conjunto despreciable el último término tiende a 0. Tomando límites en (10.6) obtenemos la fórmula del enunciado. ■

Evidentemente, este teorema es de escaso valor sin una caracterización aceptable de los conjuntos despreciables. Es claro que todo subconjunto cerrado de un conjunto despreciable para una variedad S es también despreciable.

Teorema 10.24 *Sean E y F dos subconjuntos compactos despreciables para una variedad $S \subset \mathbb{R}^n$ sin frontera. Entonces $E \cup F$ también es despreciable.*

Demostración: Sean W, $\{W_k\}_{k=1}^\infty$, $\{g_k\}_{k=1}^\infty$ según la definición de conjunto despreciable (para E) y sean W', $\{W'_k\}_{k=1}^\infty$, $\{g'_k\}_{k=1}^\infty$ los análogos para F. Basta tomar

$$W'' = W \cup W', \quad W''_k = W_k \cup W'_k, \quad g''_k = g_k g'_k.$$

Es claro que estos conjuntos y funciones prueban que $E \cup F$ es despreciable. Para la última condición observamos que

$$d(g_k g'_k) \wedge \omega = g'_k \, dg_k \wedge \omega + g_k \, dg'_k \wedge \omega.$$

■

Enunciamos el teorema siguiente en el caso en que la variedad S es un abierto en $\mathbb{R}^n$ porque es el de mayor interés en la práctica, pero afinando un poco el argumento se generaliza a abiertos en variedades arbitrarias.

Teorema 10.25 *Sea S un abierto en $\mathbb{R}^n$ y E un subconjunto compacto de $\mathbb{R}^n$ tal que*[3] *existe un cubo cerrado Q de dimensión $m \leq n-2$ y una aplicación $h : U \longrightarrow \mathbb{R}^n$ de clase C^1, donde U es un entorno de Q y $h[Q] = E$. Entonces E es despreciable para S.*

Demostración: En primer lugar observamos que podemos suponer que $m = n-2$, pues en caso contrario la aplicación f se puede componer con la proyección desde un cubo de dimensión superior. Así mismo, componiendo con una aplicación lineal podemos suponer que $Q = [0,1]^{n-2}$

Un sistema fundamental de entornos de E lo forman los conjuntos

$$W_k = \{x \in \mathbb{R}^n \mid d(x, E) < 2/k\}, \quad k = 1, 2, \ldots$$

Consideramos concretamente la distancia inducida por $\| \ \|_\infty$ en $\mathbb{R}^n$. Tomemos una función $\phi : \mathbb{R}^n \longrightarrow [0,1]$ de clase C^∞ que se anule sobre los puntos con $\|x\|_\infty \leq 1/2$ y valga 1 sobre los puntos con $\|x\|_\infty \geq 1$. Para cada natural $k > 0$ sea $\phi_k(x) = \phi(kx)$. Si C es una cota de las derivadas parciales de ϕ en $\mathbb{R}^n$, es claro que para todo $x \in \mathbb{R}^n$ se cumple $\|D_i \phi_k(x)\|_\infty \leq kC$. Observar que la cota C sólo depende de n.

Sea $I = \{l \in \mathbb{Z}^n \mid d(l/2k, E) \leq 1/k\}$. Claramente se trata de un conjunto finito. Definimos

$$g_k(x) = \prod_{l \in I} \phi_k \left(x - \frac{l}{2k} \right).$$

La función g_k es de clase C^∞. Veamos que se anula en un entorno de E, concretamente en el de los puntos $x \in \mathbb{R}^n$ tales que $d(x, E) < 1/4k$. Dado uno de estos puntos x, existe $l \in \mathbb{Z}^n$ tal que $d(x, l/2k) \leq 1/2k$ (la coordenada l_i es la parte entera de $2kx_i$). Claramente $d(l/2k, E) < 1/k$, luego $l \in I$ y $\phi_k(x - l/2k) = 0$, y en consecuencia $g_k(x) = 0$, como queríamos probar.

Veamos ahora que g_k vale 1 fuera de W_k. En efecto, si $d(x, E) \geq 2/k$ y $l \in I$, es decir, $d(l/2k, E) \leq 1/k$, entonces $d(x, l/2k) > 1/k$, luego $\phi_k(x - l/2k) = 1$ y así $g_k(x) = 1$.

El motivo de toda esta construcción es garantizar que las funciones g_k cumplen una condición adicional, y es que sus derivadas parciales están acotadas por $C_1 k$, donde C_1 es una constante que sólo depende de n. En efecto, tomemos

[3] En el caso $n = 2$ el teorema se cumple si E consta de un solo punto. Algunos razonamientos han de ser sustituidos por otros más simples.

un punto $x \in \mathbb{R}^n$ alrededor del cual las derivadas de g_k no sean idénticamente nulas, lo que implica que $\|x - l_0/2k\|_\infty \leq 1/k$ para un cierto $l_0 \in I$. De los factores que componen g_k, todos serán nulos en un entorno de x excepto a lo sumo los correspondientes a vectores $l \in I$ tales que $\|x - l/2k\|_\infty \leq 1/k$, pero entonces $\|l - l_0\|_\infty \leq 4$, y es fácil ver que hay a lo sumo 9^n puntos así. Al derivar g_k obtenemos una suma de 9^n términos, cada uno de los cuales es un producto de la derivada de una función $\phi_k(x - l/2k)$ por otras funciones de este tipo sin derivar. Éstas están acotadas por 1 y la primera por Ck, luego cada derivada de g_k está acotada por C_1k, donde C_1 es una constante que sólo depende de n.

Tomando $W = \mathbb{R}^n$ tenemos comprobada la condición a) de la definición de conjunto despreciable.

Consideremos ahora una $n-1$-forma ω de clase C^1 en $\mathbb{R}^n$. Será de la forma

$$\omega = \sum_{j=1}^n f_j\, dx_1 \wedge \cdots \wedge dx_{j-1} \wedge dx_{j+1} \wedge \cdots \wedge dx_n.$$

Entonces $dg_k \wedge \omega = f\, dx_1 \wedge \cdots \wedge dx_n$, donde la función

$$f = \sum_{j=1}^n (-1)^{j+1} f_j D_j g_k$$

está acotada por C_2k, y la constante C_2 sólo depende de n (las funciones f_j se acotan en $\overline{W}_1$). Puesto que f tiene soporte compacto, la n-forma $dg_k \wedge \omega$ es integrable en $\mathbb{R}^n$ y determina la medida dada por $\mu_k(A) = \int_A f\, dm$. Entonces

$$|\mu_k|(\mathbb{R}^n) = \int_{\mathbb{R}^n} |f|\, dm \leq C_2k\, m(W_k). \tag{10.7}$$

Ahora estimaremos la medida de W_k para concluir que la expresión anterior tiende a 0 cuando k tiende a infinito. Dividimos el cubo Q en k^{n-2} cubos de lado $1/k$. Como las normas en $\mathbb{R}^n$ son equivalentes, la distancia euclídea entre dos puntos del mismo cubo está acotada por C_3/k, para una cierta constante k. Aplicando el teorema del valor medio a cada función coordenada de h concluimos que si u y v están en el mismo cubo, entonces $\|h(u) - h(v)\|_\infty \leq C_4/k$. Si $x \in W_k$, entonces x dista menos de $2/k$ de un punto de E, el cual dista menos de C_4/k de la imagen del centro de uno de los k^{n-2} cubos, luego W_k está contenido en la unión de k^{n-2} bolas de radio C_5/k (para cualquier norma, por ejemplo la euclídea), luego

$$m(W_k) \leq k^{n-2}\frac{C_6}{k^n} = \frac{C_6}{k^2},$$

donde las constantes $C_3, \ldots, C_6$ dependen de n, f y ω, pero no de k. Conectando esto con (10.7) llegamos a que $|\mu_k|(\mathbb{R}^n) \leq C_7/k$, que tiende a 0 con k. ∎

En vista de lo anterior tenemos la versión siguiente del teorema de Stokes, que incluye como caso particular el de los cubos que ya habíamos probado:

Teorema 10.26 (Teorema de Stokes con singularidades) *Consideremos un abierto S en $\mathbb{R}^n$ tal que el conjunto de puntos singulares de su frontera*

sea unión de un número finito de variedades compactas de dimensión menor o igual que $n-2$. *Sea* ω *una* $n-1$*-forma definida en un entorno de* $\overline{S}$ *tal que la intersección con* $\overline{S}$ *de su soporte sea compacta y las formas* ω *y* $d\omega$ *sean integrables en* S *y* ∂S *respectivamente. Entonces*

$$\int_S d\omega = \int_{\partial S} \omega.$$

Evidentemente, todas las consecuencias del teorema de Stokes que hemos visto en las secciones anteriores valen ahora en el contexto de este teorema.

10.7 Apéndice: Algunas fórmulas vectoriales

Aunque el contenido de esta sección no tiene que ver directamente con el teorema de Stokes, hemos preferido posponer hasta aquí los resultados que siguen para exponerlos una vez estamos familiarizados con las principales operaciones vectoriales: gradiente, divergencia y rotacional. Los resultados principales serán las expresiones de estas operaciones en sistemas de coordenadas distintas de las cartesianas, pero antes recogemos algunas fórmulas de interés.

El operador nabla Muchas fórmulas del cálculo vectorial se recuerdan más fácilmente si definimos el "vector" nabla como

$$\nabla = \left(\frac{\partial}{\partial x}, \frac{\partial}{\partial y}, \frac{\partial}{\partial z}\right).$$

Naturalmente, esto no tiene ningún significado en sí mismo, pero formalmente el gradiente de una función f puede pensarse como el producto del vector ∇ por el escalar f, lo que concuerda con la notación ∇f. Similarmente, la divergencia de un campo vectorial V puede interpretarse como el producto escalar del vector ∇ por el vector V, lo que nos permite escribir también $\operatorname{div} V = \nabla V$. Por último, el rotacional de V es el producto vectorial de ∇ por V, o sea, $\operatorname{rot} V = \nabla \wedge V$.

En estos términos, las fórmulas siguientes quedan de forma más simétrica:

$$\nabla(fg) = g(\nabla f) + f(\nabla g) \tag{10.8}$$

$$\nabla(fV) = (\nabla f)V + f(\nabla V) \tag{10.9}$$

$$\nabla \wedge (fV) = (\nabla f \wedge V) + f(\nabla \wedge V) \tag{10.10}$$

$$\nabla(V \wedge W) = W(\nabla \wedge V) - V(\nabla \wedge W) \tag{10.11}$$

Todas ellas se comprueban sin dificultad a partir de las definiciones, aunque algunas resultan algo laboriosas.

Notar que las relaciones $\operatorname{rot}\operatorname{grad} f = 0$, $\operatorname{div}\operatorname{rot} V = 0$ también se recuerdan más fácilmente en la forma $\nabla \wedge (\nabla f) = 0$, $\nabla(\nabla \wedge V) = 0$, pues formalmente son propiedades válidas para vectores y escalares cualesquiera.

Coordenadas curvilíneas ortogonales Sea $X : U \longrightarrow S$ una aplicación de clase C^1 entre dos abiertos de $\mathbb{R}^3$. Podemos considerar a S como una variedad diferenciable y a X como una carta de S. Entonces X^{-1} es un sistema de coordenadas que a cada punto $p = (x_1, x_2, x_3) \in S$ le asigna unas coordenadas $u = (u_1, u_2, u_3)$. Supondremos que X es ortogonal, en el sentido de que los coeficientes g_{ij} del tensor métrico son nulos cuando $i \neq j$. Recordemos que $g_{ij}(u) = D_iX(u)D_jX(u)$. Abreviaremos $h_i(u) = \sqrt{g_{ii}(u)} = \|D_iX(u)\|$.

La ortogonalidad de X equivale a que los vectores $D_iX(u)$ forman en cada punto $p = X(u)$ una base ortogonal de $T_p(S)$. Por consiguiente los vectores $v_i(u) = h_i^{-1}(u)D_iX(u)$ forman una base ortonormal, a la que nos referiremos simplemente como la base asociada al sistema de coordenadas dado.

Si llamamos $u = X^{-1}$, su matriz jacobiana es $Ju(p) = JX^{-1}(u(p))$. Puesto que $(g_{ij}) = (JX)(JX)^t$ es una matriz diagonal, lo mismo le sucede a su inversa, luego $(Ju)(Ju)^t$ es diagonal en cada punto. Más concretamente:

$$\nabla u_i \nabla u_j = \begin{cases} g_{ii}^{-1} & \text{si } i = j \\ 0 & \text{si } i \neq j \end{cases}$$

De aquí podemos obtener las coordenadas de los gradientes ∇u_i en la base (v_1, v_2, v_3). En efecto, si

$$\nabla u_i = \alpha_{i1}v_1 + \alpha_{i2}v_2 + \alpha_{i3}v_3 = \frac{\alpha_{i1}}{h_1}\,D_1X + \frac{\alpha_{i2}}{h_2}\,D_2X + \frac{\alpha_{i3}}{h_3}\,D_3X,$$

aplicando la base dual du_j resulta que

$$\frac{\alpha_{ij}(p)}{h_j(p)} = du_j(p)(\nabla u_i(p)) = \nabla u_j(p)\nabla u_i(p),$$

con lo que

$$\nabla u_i = \frac{1}{h_i}\,v_i. \tag{10.12}$$

Consideremos ahora una función $f : S \longrightarrow \mathbb{R}$ de clase C^1. Llamaremos también f a su composición con X. Mediante la regla de la cadena se comprueba fácilmente que

$$\nabla f = \sum_{i=1}^{3} \frac{\partial f}{\partial u_i}\,\nabla u_i = \sum_{i=1}^{3} \frac{1}{h_i}\,\frac{\partial f}{\partial u_i}\,v_i, \tag{10.13}$$

que es la expresión del gradiente en en sistema de coordenadas considerado.

Consideremos ahora un campo vectorial en coordenadas curvilíneas

$$A = a_1(u_1, u_2, u_3)v_1 + a_2(u_1, u_2, u_3)v_2 + a_3(u_1, u_2, u_3)v_3.$$

Vamos a calcular su divergencia. Supondremos que las coordenadas están ordenadas de modo que la base (v_1, v_2, v_3) es positiva. Entonces $v_1 = v_2 \wedge v_3$, $v_2 = v_3 \wedge v_1$ y $v_3 = v_1 \wedge v_2$. Teniendo en cuenta (10.12) podemos escribir

$$A = a_1h_2h_3(\nabla u_2 \wedge \nabla u_3) + a_2h_1h_3(\nabla u_3 \wedge \nabla u_1) + a_3h_1h_2(\nabla u_1 \wedge \nabla u_2). \tag{10.14}$$

Calculamos $\operatorname{div} A$ aplicando (10.9), (10.11) y el hecho de que el rotacional de un gradiente es nulo. El resultado es

$$\nabla(a_1h_2h_3)(\nabla u_2 \wedge \nabla u_3) + \nabla(a_2h_1h_3)(\nabla u_3 \wedge \nabla u_1) + \nabla(a_3h_1h_2)(\nabla u_1 \wedge \nabla u_2).$$

Ahora aplicamos (10.13) teniendo en cuenta que un producto mixto con dos vectores iguales es nulo. Obtenemos

$$\operatorname{div} A = \left(\frac{\partial a_1h_2h_3}{\partial u_1} + \frac{\partial a_2h_1h_3}{\partial u_2} + \frac{\partial a_3h_1h_2}{\partial u_3}\right)(\nabla u_1, \nabla u_2, \nabla u_3).$$

Finalmente observamos que $(v_1, v_2, v_3) = 1$, lo que juntamente con (10.12) nos da

$$\operatorname{div} A = \frac{1}{h_1h_2h_3}\left(\frac{\partial a_1h_2h_3}{\partial u_1} + \frac{\partial a_2h_1h_3}{\partial u_2} + \frac{\partial a_3h_1h_2}{\partial u_3}\right).$$

Para calcular el rotacional de A partimos de (10.14) y aplicamos (10.10) junto con el hecho de que el rotacional de un gradiente es nulo. Así pues,

$$\operatorname{rot} A = \nabla(a_1h_1) \wedge \nabla u_1 + \nabla(a_2h_2) \wedge \nabla u_2 + \nabla(a_3h_3) \wedge \nabla u_3.$$

Aplicamos (10.13) junto con el hecho de que el rotacional de un gradiente es nulo.

$$\begin{aligned}
\operatorname{rot} A &= \frac{\partial a_1h_1}{\partial u_2}\nabla u_2 \wedge \nabla u_1 + \frac{\partial a_1h_1}{\partial u_3}\nabla u_3 \wedge \nabla u_1 \\
&= \frac{\partial a_2h_2}{\partial u_1}\nabla u_1 \wedge \nabla u_2 + \frac{\partial a_2h_2}{\partial u_3}\nabla u_3 \wedge \nabla u_2 \\
&= \frac{\partial a_3h_3}{\partial u_1}\nabla u_1 \wedge \nabla u_3 + \frac{\partial a_3h_3}{\partial u_2}\nabla u_2 \wedge \nabla u_3.
\end{aligned}$$

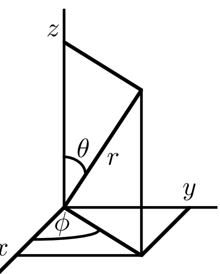
Por último aplicamos (10.12) y agrupamos los coeficientes de cada v_i. El resultado se recuerda mejor mediante la regla mnemotécnica

$$\operatorname{rot} A = \frac{1}{h_1h_2h_3}\begin{vmatrix} h_1v_1 & h_2v_2 & h_3v_3 \\ \frac{\partial}{\partial u_1} & \frac{\partial}{\partial u_2} & \frac{\partial}{\partial u_3} \\ h_1a_1 & h_2a_2 & h_3a_3 \end{vmatrix}.$$

La relación $\Delta f = \operatorname{div} \nabla f$ nos da inmediatamente la expresión del laplaciano:

$$\Delta f = \frac{1}{h_1h_2h_3}\left(\frac{\partial}{\partial u_1}\left(\frac{h_2h_3}{h_1}\frac{\partial f}{\partial u_1}\right) + \frac{\partial}{\partial u_2}\left(\frac{h_1h_3}{h_2}\frac{\partial f}{\partial u_2}\right) + \frac{\partial}{\partial u_3}\left(\frac{h_1h_2}{h_3}\frac{\partial f}{\partial u_3}\right)\right).$$

Notemos que las fórmulas anteriores se reducen a las usuales en el caso de las coordenadas cartesianas, para las cuales $h_1 = h_2 = h_3 = 1$. Existen varios sistemas de coordenadas que ayudan con frecuencia en los cálculos con vectores.



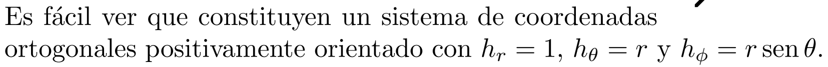
Citaremos por ejemplo el caso de las *coordenadas esféricas,* definidas sobre el abierto

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 \neq 0\}.$$

En un entorno de cada punto vienen dadas por

$$X(r, \theta, \phi) = (r \operatorname{sen} \theta \cos \phi, r \operatorname{sen} \theta \operatorname{sen} \phi, r \cos \theta).$$

Es fácil ver que constituyen un sistema de coordenadas


ortogonales positivamente orientado con $h_r = 1$, $h_\theta = r$ y $h_\phi = r \operatorname{sen} \theta$.

Otro caso es el de las *coordenadas cilíndricas,* definidas en el mismo abierto y dadas por

$$X(r, \theta, z) = (r \cos \theta, r \operatorname{sen} \theta, z).$$

Claramente $h_r = 1$, $h_\theta = r$, $h_z = 1$.

Capítulo XI

Cohomología de De Rham

La regla de Barrow reduce el problema de calcular la integral de una función continua cualquiera al cálculo de una primitiva. En la práctica existen funciones para las cuales la determinación de una primitiva puede ser muy complicado, e incluso puede ocurrir que dicha primitiva no admita una expresión en términos de funciones "conocidas", como senos, exponenciales, etc. y que por consiguiente la citada regla, aunque pueda ser de utilidad, no resuelve completamente el problema. No obstante, contamos con el hecho de que toda función continua en un intervalo tiene primitiva o, lo que es lo mismo, que toda 1-forma continua $f\,dx$ en un intervalo puede expresarse como dg, para una cierta 0-forma g.

La situación es distinta en el caso general: no es cierto que toda k-forma ω sea la diferencial de una $k-1$-forma, ni siquiera en el caso de 1-formas sobre 1-variedades distintas de los intervalos. El teorema de Stokes hace que sea interesante determinar bajo qué condiciones una forma es la diferencial de otra. Además, bajo este planteamiento caben problemas muy variados. Por ejemplo, un campo F en $\mathbb{R}^n$ es conservativo si y sólo si es el gradiente de una función, lo cual equivale a que la 1-forma $F\,d\vec{r}$ sea la diferencial de una 0-forma.

Definición 11.1 Diremos que una k-forma ω es *exacta* si $\omega = d\omega'$, para una cierta $k-1$-forma ω'. Diremos que ω es *cerrada* si $d\omega = 0$.

Es obvio que una condición necesaria para que una forma sea exacta es que sea cerrada, pues si $\omega = d\omega'$ entonces $d\omega = d(d\omega')) = 0$. Sucede que en muchos casos esta simple condición es también suficiente. En este capítulo veremos lo más básico de una importante teoría que nos permite determinar bajo qué condiciones esto es así.

11.1 Grupos de cohomología

Nota *En las secciones siguientes, cuando hablemos de variedades diferenciables entenderemos que tienen cartas de clase C^∞ y cuando digamos que una función o una forma es diferenciable entenderemos que es de clase C^∞.*

Comenzamos con unas definiciones algebraicas que que se ajustan a la estructura de las álgebras de Grassmann:

Definición 11.2 Un *módulo graduado* sobre un anillo A es una suma directa de A-módulos $C = \bigoplus\limits_{k\in\mathbb{Z}} C_k$.

Los elementos de cada submódulo C_k se llaman *homogéneos* de grado k.

Un *submódulo graduado* de C es un módulo graduado D tal que $D_k = C_k \cap D$.

Un *homomorfismo graduado* $f : C \longrightarrow D$ (de grado d) entre módulos graduados es un homomorfismo de módulos tal que $f_k = f|_{C_k} : C_k \longrightarrow D_{k+d}$ para todo entero k.

Un *complejo* es un par ordenado $\mathcal{C} = (C, \partial)$, donde C es un módulo graduado y $\partial : C \longrightarrow C$ es un homomorfismo de grado ± 1 tal que $\partial \circ \partial = 0$. Si ∂ tiene grado -1 el complejo se dice *directo* y si el grado es 1 se dice *inverso.*

Un complejo directo puede verse también como una sucesión de módulos y homomorfismos:

$$\cdots \xrightarrow{\partial_{k+2}} C_{k+1} \xrightarrow{\partial_{k+1}} C_k \xrightarrow{\partial_k} C_{k-1} \xrightarrow{\partial_{k-1}} \cdots$$

de modo que $\partial_{k+1} \circ \partial_k = 0$ para todo k. Un complejo inverso es igual pero cambiando el sentido de las flechas.

Ejemplo Si S es una variedad diferenciable, el álgebra de Grassmann

$$\Lambda(S) = \bigoplus_{k\in\mathbb{Z}} \Lambda^k(S)$$

es un espacio vectorial graduado (tomando $\Lambda^k(S) = 0$ para $k < 0$). Además la diferencial exterior d es un homomorfismo de grado 1 en ΛS que lo convierte en un complejo inverso. ■

Notemos que todo complejo directo puede verse como un complejo inverso sin más que cambiar los índices, por lo que todo lo que vale para complejos directos vale para complejos inversos. La diferencia la marca únicamente la práctica: del mismo modo que resultaría artificial tratar al álgebra de Grassmann como un complejo directo, hay otros complejos que resultaría artificial tratarlos como complejos inversos.

Es costumbre usar términos distintos para referirse a los conceptos correspondientes a los complejos directos y a sus análogos en los complejos inversos. Para empezar, los módulos homogéneos de un complejo inverso se suelen representar con superíndices C^k en lugar de con subíndices C_k.

Dado un complejo $\mathcal{C} = (C, \partial)$, el homomorfismo ∂ recibe el nombre de *operador frontera* en un complejo directo y *operador cofrontera* en un complejo inverso.

Los elementos de C_k (en un complejo directo) se llaman *cadenas* de dimensión k y los de C^k (en un complejo inverso) se llaman *cocadenas* de dimensión k.

Los elementos de $Z_k = \mathrm{N}(\partial_k)$ (el núcleo de ∂_k) se llaman *ciclos* de dimensión k. Respectivamente, los elementos de $Z^k = \mathrm{N}(\partial^k)$ se llaman *cociclos* de dimensión k.

Los elementos de $F_k = \mathrm{Im}(\partial_{k+1})$ (resp. $F^k = \mathrm{Im}(\partial^{k-1})$) se llaman *fronteras* (resp. *cofronteras*) de dimensión k.

La condición $\partial \circ \partial = 0$ implica que $F_k \subset Z_k$ (resp. $F^k \subset Z^k$). El módulo cociente $H_k(\mathcal{C}) = Z_k/F_k$ (resp. $H^k(\mathcal{C}) = Z^k/F^k$) recibe el nombre de *grupo de homología* (resp. *grupo de cohomología*) de dimensión k. Dos (co)ciclos son *(co)homólogos* si pertenecen al la misma clase de (co)homología.

Una vez entendida la cuestión de notación, en lo que sigue desarrollaremos la teoría en términos de cohomología, pues las álgebras de Grassmann son complejos inversos. Si $\mathcal{C}$ es un complejo, definimos

$$H(\mathcal{C}) = \bigoplus_{k\in\mathbb{Z}} H^k(\mathcal{C}),$$

que es obviamente un módulo graduado.

Definición 11.3 Sea S una variedad diferenciable. El grupo de cohomología de dimensión k del álgebra de Grassmann $\Lambda(S)$ recibe el nombre de *grupo de cohomología de De Rham* de dimensión k de la variedad S y se representa por $H^k(S)$. Llamaremos

$$H(S) = \bigoplus_{k\in\mathbb{Z}} H^k(S).$$

Notemos que las cocadenas de dimensión k son las k-formas, los cociclos son las k-formas cerradas y las cofronteras son las k-formas exactas. Los grupos de cohomología son en este caso espacios vectoriales sobre $\mathbb{R}$.

Si S es una variedad de dimensión n es evidente que $H^k(S) = 0$ para $k < 0$ y $k > n$. Los 0-cociclos son las funciones en S cuya diferencial es nula. Si S es conexa son exactamente las funciones constantes y como $F^0 = 0$, concluimos que $H^0(S) \cong \mathbb{R}$. Más en general, es fácil ver que si S tiene p componentes conexas entonces $H^0(S) \cong \mathbb{R}^p$.

El objetivo de la teoría que estamos desarrollando es calcular los grupos de cohomología $H^k(S)$ para $1 \leq k \leq n$, pues si probamos que $H^k(S) = 0$ entonces sabemos que las k-formas exactas coinciden con las cerradas y tenemos así una caracterización sencilla de las primeras.

Definición 11.4 Un *homomorfismo de complejos* $\phi : \mathcal{C} \longrightarrow \mathcal{C}'$ es un homomorfismo de grado 0 tal que $\phi \circ \partial' = \partial \circ \phi$, o equivalentemente, tal que los diagramas

siguientes conmutan:

$$\begin{array}{ccccccccc}
\cdots \longrightarrow & C^{k-1} & \xrightarrow{\partial^{k-1}} & C^k & \xrightarrow{\partial^k} & C^{k+1} & \longrightarrow \cdots \\
 & \downarrow \phi^{k-1} & & \downarrow \phi^k & & \downarrow \phi^{k+1} & \\
\cdots \longrightarrow & C'^{k-1} & \xrightarrow{\partial'^{k-1}} & C'^k & \xrightarrow{\partial'^k} & C'^{k+1} & \longrightarrow \cdots
\end{array}$$

Es claro que un homomorfismo ϕ envía ciclos a ciclos y fronteras a fronteras, luego induce homomorfismos $\overline{\phi^k} : H^k(\mathcal{C}) \longrightarrow H^k(\mathcal{C}')$ o, equivalentemente, induce un homomorfismo de grado 0

$$\overline{\phi} : H(\mathcal{C}) \longrightarrow H(\mathcal{C}').$$

Es inmediato que la composición de homomorfismos de complejos es un homomorfismo de complejos así como que $\overline{\phi \circ \psi} = \overline{\phi} \circ \overline{\psi}$. Si ϕ es un isomorfismo entonces $\overline{\phi}$ también lo es y $(\overline{\phi})^{-1} = \overline{\phi^{-1}}$. Si $f : S \longrightarrow T$ es una aplicación diferenciable entre variedades, la retracción $f^\sharp : \Lambda(T) \longrightarrow \Lambda(S)$ es un homomorfismo de complejos, que a su vez induce un homomorfismo $\overline{f^\sharp} : H(T) \longrightarrow H(S)$. Por simplificar la notación escribiremos $\overline{f}$ en lugar de $\overline{f^\sharp}$.

Ejercicio: Sea $S = \bigcup_{i\in I} S_i$ una unión disjunta de variedades (entendiendo que cada una de ellas es abierta y cerrada en S). Probar que $H(S) \cong \prod_{i\in I} H(S_i)$.

Los difeomorfismos entre variedades inducen isomorfismos entre los grupos de cohomología de De Rham. Esto era de esperar. Sin embargo, vamos a probar que hay aplicaciones mucho más generales que los difeomorfismos y que también inducen isomorfismos entre los grupos de cohomología, lo que nos permitirá reducir el cálculo de unas variedades a otras más simples. Dedicamos a ello la sección siguiente.

11.2 Homotopías

Recordemos que tenemos definido el producto de una variedad sin frontera por una variedad con o sin frontera. Claramente, el producto de una recta y una circunferencia nos da una superficie cilíndrica, mientras que el producto de una recta y un círculo nos da un cilindro sólido (con o sin frontera, según si el círculo es cerrado o abierto). En general, dada una variedad S, llamaremos *cilindro* de S a la variedad producto $\mathbb{R} \times S$. Conviene pensar en $\mathbb{R} \times S$ como en infinitas copias de S "apiladas" una encima de otra, cada una a una altura t. Concretamente, la copia de altura $t \in \mathbb{R}$ es $S_t = \{t\} \times S$. Es claro que S_t es una variedad difeomorfa a S.

Cuando trabajemos con un cilindro $\mathbb{R} \times S$ consideraremos únicamente cartas de la forma $\bar{X} = I \times X$, donde I es la identidad en $\mathbb{R}$ y X es una carta en S. De este modo, $\bar{X}(t, u) = \big(t, X(u)\big)$. En particular vemos que el primer vector de la base canónica está en todos los espacios tangentes:

$$e_1 = D_1\bar{X}(t, u) \in T_{(t,p)}(\mathbb{R} \times S).$$

Definición 11.5 Dos aplicaciones $f, g : S_1 \longrightarrow S_2$ diferenciables entre dos variedades son *homotópicas* si existe $H : \mathbb{R} \times S_1 \longrightarrow S_2$ diferenciable de modo que para todo $p \in S_1$ se cumpla

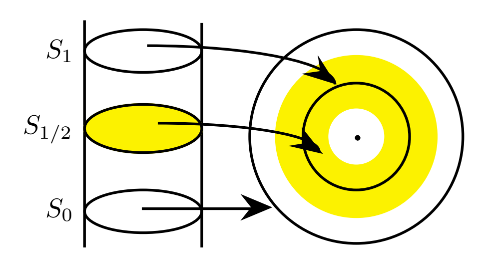
$$H(0, p) = f(p), \quad H(1, p) = g(p).$$

Se dice que la aplicación H es una *homotopía* entre f y g.

Es decir, dos aplicaciones son homotópicas si una se puede transformar en la otra mediante una gradación diferenciable.

Ejemplo Sea S_1 la bola abierta de centro 0 y radio 2 en $\mathbb{R}^n$ menos su centro 0. Se trata de un abierto en $\mathbb{R}^n$ y por lo tanto es una variedad. Sea $f : S_1 \longrightarrow S_1$ la aplicación identidad y sea $g : S_1 \longrightarrow S_1$ la aplicación dada por $g(x) = x/\|x\|$. Ambas son homotópicas. Basta considerar la homotopía

$$H(t, x) = \frac{tx}{\|x\|} + (1 - t)x.$$



Si fijamos x y hacemos variar t entre 0 y 1 vemos que $H(t, x)$ recorre el segmento radial que va desde x hasta la circunferencia unidad. Así, H transforma S_0 en todo S_0 (es la identidad), al llegar a $1/2$ cada punto ha recorrido la mitad de su camino hasta la circunferencia, por lo que la imagen de $S_{1/2}$ es la corona sombreada en la figura, y finalmente la imagen de S_1 es la circunferencia. La homotopía H aproxima paulatinamente la imagen de cada punto por la aplicación f hasta su imagen por la aplicación g. ■

El objetivo de esta sección es probar que dos aplicaciones homotópicas entre variedades inducen la misma aplicación entre los grupos de cohomología. La prueba es una generalización de un teorema de Poincaré y necesita varios conceptos previos.

La evaluación Sea $S \subset \mathbb{R}^m$ una variedad y V un campo de vectores tangentes en S, es decir, $V : S \longrightarrow \mathbb{R}^m$ es una función diferenciable y para cada $p \in S$ se cumple $V(p) \in T_p(S)$. Entonces V induce una aplicación lineal de grado -1 $i(V) : \Lambda(S) \longrightarrow \Lambda(S)$ que a cada k-forma ω le asigna la $k-1$-forma dada por

$$i(V)(\omega)(p)(v_1, \ldots, v_{k-1}) = \omega(p)(V(p), v_1, \ldots, v_{k-1}).$$

Convenimos que $i(V)(f) = 0$ para toda $f \in \Lambda^0(S)$.

Es claro que $i(V)(\omega)(p) \in A^{k-1}(T_p(S))$, pero falta ver que $i(V)(\omega)$ es diferenciable. En principio $i(V)$ es una aplicación lineal de $\Lambda(S)$ en el álgebra

de todas las formas en S, no necesariamente diferenciables. Una comprobación rutinaria nos da que si $\omega_1 \in \Lambda^k(S)$, $\omega_2 \in \Lambda(S)$, entonces[1]

$$i(V)(\omega_1 \wedge \omega_2) = i(V)(\omega_1) \wedge \omega_2 + (-1)^k \omega_1 \wedge i(V)(\omega_2).$$

Las aplicaciones lineales que cumplen esta relación se llaman *antiderivaciones*. Otro ejemplo de antiderivación es la diferencial exterior.

Para probar que $i(V)(\omega)$ es diferenciable en un punto p tomamos una carta X alrededor de p. Si W es el rango de X, notamos que $i(V|_W)(\omega|_W)$ coincide con $i(V)(\omega)|_W$, luego basta probar que la primera es diferenciable. Ahora bien, una forma en W se expresa en función del producto exterior a partir de 0-formas y las diferenciales de las coordenadas dx_i y al ser una antiderivación $i(V|_W)(\omega|_W)$ quedará en función de las imágenes por $i(V|_W)$ de estas formas en particular, luego basta ver que $i(V|_W)(dx_i)$ es diferenciable (para las 0-formas es obvio, porque la imagen es nula). Ahora bien,

$$i(V|_W)(dx_i)(p) = dx_i(p)(V(p)) = V_i(p),$$

donde $V(X(x)) = \sum_{i=1}^{n} V_i(X(x))\, D_i X(x)$. El teorema de la función implícita justifica que las funciones $X \circ V_i$ son diferenciables (luego las V_i también).

La antiderivación $i(V)$ se llama *evaluación* en V.

En particular, en un cilindro $\mathbb{R} \times S$ podemos considerar el campo constante igual a e_1. Si S se puede cubrir con una sola carta X y llamamos $(t, x_1, \ldots, x_n)$ a las coordenadas del cilindro, tenemos que $i(e_1)(dt) = 1$, $i(e_1)(dx_i) = 0$. Estas relaciones determinan a $i(e_1)$. Concretamente:

$$i(e_1)(f\, dt \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k}) = f\, dx_{i_1} \wedge \cdots \wedge dx_{i_k},$$

$$i(e_1)(f\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}) = 0.$$

Las inclusiones Consideremos ahora las inclusiones $j_t : S \longrightarrow \mathbb{R} \times S$ dadas por $j_t(p) = (t, p)$. Obviamente son diferenciables, por lo que tienen asociadas las retracciones $j_t^\sharp : \Lambda(\mathbb{R} \times S) \longrightarrow \Lambda(S)$, que son homomorfismos de grado 0. Si S se puede cubrir por una sola carta tenemos

$$j_t^\sharp(f) = j_t \circ f, \quad j_t^\sharp(dt) = 0, \quad j_t^\sharp(dx_i) = dx_i,$$

con lo que quedan completamente determinadas.

[1]En la definición de $\omega_1 \wedge \omega_2$ separamos los sumandos correspondientes a las permutaciones que dejan a $V(p)$ entre las k primeras componentes y las que lo dejan entre las siguientes. En el primer sumando sustituimos cada permutación σ por permutación $\bar{\sigma}$ que resulta de llevar $V(p)$ a la primera posición. El cambio de signo que sufre ω se compensa con el cambio de signatura de σ a $\bar{\sigma}$. Así tenemos una suma sobre las permutaciones $\bar{\sigma}$ tales que $\bar{\sigma}(1) = 1$. Identificándolas con las permutaciones de $k + k' - 1$ elementos obtenemos $i(V)(\omega_1)(p) \wedge \omega_2(p)$. Con el segundo sumando razonamos igual, salvo que llevamos $V(p)$ a la posición $k+1$. Ahora, al pasar de las permutaciones que cumplen $\bar{\sigma}(1) = k + 1$ a la permutación correspondiente de $k + k' - 1$ elementos la signatura varía en $(-1)^k$.

El operador integral Finalmente, dados dos números reales $a < b$, definimos una aplicación lineal $I_a^b : \Lambda(\mathbb{R}\times S) \longrightarrow \Lambda(S)$ de grado 0 que a cada $\omega \in \Lambda^k(\mathbb{R}\times S)$ le asigna

$$I_a^b(\omega)(p)(v_1,\ldots,v_k) = \int_a^b j_t^\sharp(\omega)(p)(v_1,\ldots,v_k)\,dt.$$

Fijado un punto $p \in S$ y una carta X a su alrededor, la restricción de ω al rango de $I \times X$ se expresa como suma de k-formas de tipo

$$\eta = f\,dx_{i_1} \wedge \cdots \wedge dx_{i_k},$$

donde admitimos la posibilidad de que $i_1 = 0$ con el convenio de que $x_0 = t$. Si aparece dt, entonces $j_t^\sharp(\eta) = 0$ y el término no contribuye en nada. En caso contrario

$$j_t^\sharp(\eta)(p)(v_1,\ldots,v_k) = f(t,p)(dx_{i_1} \wedge \cdots \wedge dx_{i_k})(v_1,\ldots,v_k),$$

luego

$$I_a^b(\eta)(p)(v_1,\ldots,v_k) = \left(\int_a^b f(t,p)\,dt\right)(dx_{i_1} \wedge \cdots \wedge dx_{i_k})(v_1,\ldots,v_k),$$

y en definitiva

$$I_a^b(\eta) = \left(\int_a^b f\,dt\right) dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

El teorema 7.23 implica que $I_a^b(\eta)$ es una forma diferenciable. En general $I_a^b(\omega)$ (restringida al rango de $I \times X$) es una suma de formas de este tipo, luego efectivamente $I_a^b(\omega) \in \Lambda(S)$. Evidentemente I_a^b es lineal. Veamos que conmuta con la diferencial exterior, es decir: $d \circ I_a^b = I_a^b \circ d$.

Sea $p \in S$ y X una carta alrededor de p. Es claro que el valor que toma en p la imagen de una forma por cualquiera de los dos miembros será la misma que la imagen de la restricción de dicha forma al rango de la carta $I \times X$ por el operador integral en dicho abierto. Así pues, podemos trabajar con una forma definida en el rango de $I \times X$. Como ambos miembros son lineales podemos tomarla de tipo

$$\omega = f\,dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

Si $i_1 = 0$, es decir, si ω contiene a dt, entonces $d\omega$ se expresará como suma de formas, todas ellas con dt, luego tanto si hacemos actuar primero la diferencial como el operador integral obtenemos la forma nula. Supongamos, pues, que ω no contiene a dt. Entonces, tanto en un orden como en otro, llegamos a

$$\sum_{i \neq i_j} \left(\int_a^b \frac{\partial f}{\partial x_i}\,dt\right) dx_i \wedge dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

Ahora probamos una igualdad que relaciona todas las funciones que acabamos de introducir y a partir de la cual se deducirá fácilmente el resultado que queremos probar sobre homotopías. Veamos que

$$j_b^\sharp - j_a^\sharp = d \circ i(e_1) \circ I_a^b + i(e_1) \circ I_a^b \circ d. \qquad (11.1)$$

Puesto que el operador integral conmuta con la diferencial, podemos escribir el segundo miembro como $(d \circ i(e_1) + i(e_1) \circ d) \circ I_a^b$. Por el argumento habitual podemos restringirnos al rango de una carta y trabajar con una forma

$$\omega = f\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

A su vez hemos de distinguir si aparece dt o no. Si no aparece tenemos que $d(i(e_1)(\omega)) = 0$ y

$$i(e_1)(d\omega) = \frac{\partial f}{\partial t}\, dx_{i_1} \wedge \cdots \wedge dx_{i_k}.$$

Al aplicar I_a^b obtenemos

$$\left(\int_a^b \frac{\partial f}{\partial t}\, dt\right) dx_{i_1} \wedge \cdots \wedge dx_{i_k} = (j_b \circ f - f_a \circ f) dx_{i_1} \wedge \cdots \wedge dx_{i_k} = j_b^\sharp(\omega) - j_a^\sharp(\omega).$$

Supongamos ahora que $\omega = f\, dt \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k}$. Entonces

$$d\omega = -\sum_{i \neq i_j} \frac{\partial f}{\partial x_i}\, dt \wedge dx_i \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k},$$

luego

$$i(e_1)(d\omega) = -\sum_{i \neq i_j} \frac{\partial f}{\partial x_i}\, dx_i \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k}$$

y

$$d(i(e_1)(\omega)) = \sum_{i \neq i_j} \frac{\partial f}{\partial x_i}\, dx_i \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k} + \frac{\partial f}{\partial t}\, dt \wedge dx_{i_2} \wedge \cdots \wedge dx_{i_k}.$$

Al sumar estos dos términos nos queda sólo el último sumando de la última igualdad y, como tiene dt, al aplicar I_a^b queda la forma nula. Así mismo es claro que $j_b^\sharp(\omega) - j_a^\sharp(\omega) = 0$. ■

Ahora conviene introducir el concepto siguiente:

Definición 11.6 Sean $\phi, \psi : \mathcal{C} \longrightarrow \mathcal{C}'$ homomorfismos de complejos inversos. Diremos que son *homotópicos* si existe un homomorfismo $h : \mathcal{C} \longrightarrow \mathcal{C}'$ de grado -1 tal que $\phi - \psi = h\partial' + \partial h$. Equivalentemente, tal que

$$\phi^k - \psi^k = \partial^k h^{k+1} + h^k \partial^{k-1}.$$

En tal caso diremos que h es una *homotopía*[2] entre ϕ y ψ.

[2]Las homotopías entre homomorfismos de complejos directos tienen grado 1

Es evidente que si ϕ y ψ son homotópicos entonces $\psi - \phi$ transforma cociclos es cofronteras, por lo que $\overline{\phi} = \overline{\psi}$.

Casi tenemos probado el teorema siguiente:

Teorema 11.7 *Si $f, g : S_1 \longrightarrow S_2$ son aplicaciones homotópicas entre variedades, entonces $f^\sharp, g^\sharp : \Lambda(S_2) \longrightarrow \Lambda(S_1)$ son homomorfismos homotópicos.*

DEMOSTRACIÓN: Sea $H : \mathbb{R} \times S_1 \longrightarrow S_2$ una homotopía entre f y g. Definimos $h = H^\sharp \circ i(e_1) \circ I_0^1$. Claramente $h : \Lambda(S_2) \longrightarrow \Lambda(S_1)$ es un homomorfismo de grado -1. Componiendo con $H^\sharp$ en ambos miembros de (11.1) obtenemos

$$H^\sharp \circ (j_1^\sharp - j_0^\sharp) = H^\sharp \circ d \circ i(e_1) \circ I_0^1 + H^\sharp \circ i(e_1) \circ I_0^1 \circ d.$$

El primer miembro es $(j_1 \circ H)^\sharp - (j_0 \circ H)^\sharp = g^\sharp - f^\sharp$. Teniendo en cuenta que $H^\sharp$ conmuta con la diferencial, el segundo miembro es $d \circ h + h \circ d$, luego h es una homotopía. ■

Definición 11.8 Sean $S_1 \subset S_2$ variedades diferenciables. Una *retracción* de S_2 en S_1 es una aplicación $r : S_2 \longrightarrow S_1$ diferenciable tal que $r|_{S_1}$ sea la identidad. En tal caso se dice que S_1 es un *retracto* de S_2. Si la retracción es homotópica a la identidad en S_2 entonces se dice que la variedad S_1 es un *retracto por deformación*[3] de S_2.

Informalmente, S_1 es un retracto por deformación de S_2 si S_2 puede deformarse gradualmente hasta quedar aplastado sobre S_1 y ello sin mover ninguno de los puntos de S_1.

Por ejemplo, la esfera unitaria de dimensión n es un retracto de la bola (abierta o cerrada) de dimensión $n+1$ de centro 0 y radio 2 menos el origen, tal y como muestra el ejemplo que hemos visto de homotopía. En realidad es claro que los centros y los radios son irrelevantes: cualquier bola menos su centro se puede retraer homotópicamente hasta cualquier esfera concéntrica contenida en ella. La deformación consiste en agrandar paulatinamente el agujero que deja el centro y contraer los puntos exteriores a la esfera.

Una superficie cilíndrica puede retraerse hasta una circunferencia (sin más que aplastar verticalmente sus paredes). Un toro sólido puede "estrangularse" hasta una circunferencia.

Las variedades que pueden retraerse a un punto se llaman *contractibles.* Entre ellas se encuentran $\mathbb{R}^n$, las bolas abiertas y cerradas y, más en general, todas las variedades cubribles por una sola carta con dominio contractible. En cambio, una esfera no es contractible (como veremos enseguida).

El interés de todo esto radica en que los retractos por deformación tienen la misma cohomología:

[3]Estos conceptos tienen interés para espacios vectoriales arbitrarios, considerando entonces retracciones y homotopías continuas, ya no diferenciables.

Teorema 11.9 *Sea S_1 un retracto por deformación de una variedad S_2. Entonces la inclusión $i : S_1 \longrightarrow S_2$ induce un isomorfismo $\bar{i} : H(S_2) \longrightarrow H(S_1)$.*

Demostración: Sea $r : S_2 \longrightarrow S_1$ una retracción homotópica a la identidad en S_2. Entonces $i \circ r = I|_{S_1}$, luego $\bar{r} \circ \bar{i} = I|_{H(S_1)}$. Por otra parte, $r \circ i$, es decir, r vista como aplicación de S_2 en S_2, es homotópica a la identidad, luego el teorema anterior nos da que $\overline{r \circ i} = I|_{H(S_2)}$, luego $\bar{i} \circ \bar{r} = I|_{H(S_2)}$.

Estas relaciones prueban que $\bar{i}$ y $\bar{r}$ son biyectivas y mutuamente inversas. ■

Así pues, si queremos conocer la cohomología de todas las variedades contractibles no tenemos más que estudiar la más simple de todas: el punto.

Teorema 11.10 *Sea S una variedad contractible. Entonces*

$$H^0(S) \cong \mathbb{R}, \quad H^k(S) = 0 \quad \textit{para } k \neq 0.$$

Demostración: Según lo dicho basta estudiar la cohomología de De Rham de la 0-variedad S formada por un punto. La variedad tangente es trivial, luego $\Lambda^k(S) = 0$ para $k > 0$ y consecuentemente todos los grupos de cohomología (salvo el primero) son triviales.[4] ■

Para terminar la sección observamos que, tal y como habíamos afirmado, las esferas no son contractibles. Más en general, ninguna variedad compacta orientable S sin frontera es contractible. En efecto, si S tiene dimensión n, el teorema de Stokes implica que la diferencial de una $n-1$-forma en S tiene integral nula, pero el elemento de medida de S es una n-forma cuya integral no es nula, luego no es la diferencial de ninguna $n-1$-forma, y obviamente es cerrada, luego $H^n(S) \neq 0$.

11.3 Sucesiones exactas

Ya hemos visto cómo las homotopías nos permiten relacionar los grupos de cohomología de variedades distintas. Otra potente herramienta para relacionar grupos de cohomología son las sucesiones exactas:

Definición 11.11 Diremos que una sucesión de homomorfismos de módulos

$$\cdots \xrightarrow{\phi_{k-2}} M_{k-1} \xrightarrow{\phi_{k-1}} M_k \xrightarrow{\phi_k} M_{k+1} \xrightarrow{\phi_{k+1}} \cdots$$

es *exacta* en M_k si $\operatorname{Im} \phi_{k-1} = \mathrm{N}(\phi_k)$. Diremos que es *exacta* si lo es en todos los módulos.

[4] Quizá el lector ponga en duda que los teoremas que hemos probado pensando en variedades arbitrarias sean aplicables a un punto. Lo cierto es que así es, pero, de todos modos, si S es un espacio contractible entonces la identidad es homotópica a una función constante c, pero $\bar{I} = I|_{H(S)}$ y $\overline{c^k} = 0$ para $k > 0$, todo ello sin hacer referencia a 0-variedades.

Notemos que la exactitud en M de una sucesión de la forma

$$N \xrightarrow{\alpha} M \longrightarrow 0$$

equivale a que α sea suprayectiva (en situaciones como ésta se entiende que la flecha sin nombre representa al homomorfismo nulo, pues no hay otra posibilidad). Igualmente, la exactitud en M de una sucesión

$$0 \longrightarrow M \xrightarrow{\alpha} N$$

equivale a que α sea inyectiva. Por consiguiente, una sucesión exacta de la forma

$$0 \longrightarrow M \longrightarrow N \longrightarrow 0$$

nos da que M y N son isomorfos. Las situaciones de este tipo son las que hacen útiles a las sucesiones exactas. El resultado principal de esta sección es un teorema que a partir de una sucesión exacta entre complejos nos permite construir una sucesión exacta entre sus grupos de cohomología. Como aplicación obtendremos la cohomología de las esferas. Veamos primero un resultado auxiliar.

Teorema 11.12 *Consideremos el siguiente diagrama conmutativo de módulos y supongamos que sus filas son exactas.*

$$\begin{array}{ccccccc} & & Z'^1 & \xrightarrow{\phi'} & Z'^2 & \xrightarrow{\psi'} & Z'^3 \longrightarrow 0 \\ & & \downarrow \partial^1 & & \downarrow \partial^2 & & \downarrow \partial^3 \\ 0 \longrightarrow & & Z^1 & \xrightarrow[\phi]{} & Z^2 & \xrightarrow[\psi]{} & Z^3 \end{array}$$

Entonces existe un homomorfismo de módulos $\partial^* : \mathrm{N}(\partial^3) \longrightarrow Z^1/\operatorname{Im}\partial^1$ *tal que la sucesión*

$$N(\partial^1) \xrightarrow{\phi''} \mathrm{N}(\partial^2) \xrightarrow{\psi''} \mathrm{N}(\partial^3) \xrightarrow{\delta^*} Z^1/\operatorname{Im}\partial^1 \xrightarrow{\bar\phi} Z^2/\operatorname{Im}\partial^2 \xrightarrow{\bar\psi} Z^3/\operatorname{Im}\partial^3$$

es exacta, donde ϕ'' *y* ψ'' *son las restricciones de* ϕ' *y* ψ' *a* $\mathrm{N}(\partial^1)$ *y* $\mathrm{N}(\partial^2)$ *y* $\bar\phi$, $\bar\psi$ *son los homomorfismos inducidos de forma natural.*

Demostración: Es fácil comprobar que las aplicaciones ϕ'', ψ'', $\bar\phi$ y $\bar\psi$ están bien definidas, así como la exactitud de la sucesión en $\mathrm{N}(\partial^2)$ y $Z^2/\operatorname{Im}\partial^2$.

Para definir δ^* tomamos $c_3' \in \mathrm{N}(\partial^3)$. Entonces existe $c_2' \in Z'^2$ tal que $c_3' = \psi'(c_2')$. Como $\psi(\partial^2(c_2')) = \partial^3(\psi'(c_2')) = \partial^3(c_3') = 0$, existe un $c_1 \in Z^1$ tal que $\phi(c_1) = \partial^2(c_2')$.

Es claro que c_2' es único módulo $\mathrm{N}(\psi') = \operatorname{Im}\phi'$, luego $\partial^2(c_2')$ es único módulo $\phi[\operatorname{Im}\partial^1]$, luego c_1 es único módulo $\operatorname{Im}\partial^1$.

Por lo tanto podemos definir $\delta^*(c_3') = c_1 + \operatorname{Im}\partial^1$. Es claro que, así definido, es un homomorfismo de módulos. (Observar que en definitiva δ^* se calcula eligiendo una antiimagen por ψ', su imagen por ∂^2 y una antiimagen por ψ.)

Es claro que $\operatorname{Im}\psi'' \subset \mathrm{N}(\delta^*)$. Si $c_3' \in \mathrm{N}(\delta^*)$ entonces $c_1 = \partial^1(c_1')$, para un cierto $c_1' \in Z_1'$, luego $\partial^2(c_2') = \phi(c_1) = \phi(\partial^1(c_1')) = \partial^2(\phi'(c_1'))$, con lo

que $c_2' - \phi'(c_1') \in \mathrm{N}(\partial^2)$ y así $c_3' = \psi'(c_2') = \psi'\big(c_2' - \phi'(c_1')\big) + \psi'\big(\phi'(c_1')\big) = \psi'\big(c_2' - \phi'(c_1')\big) \in \operatorname{Im}\psi''$.

También es claro que $\operatorname{Im}\delta^* \subset \mathrm{N}(\bar{\phi})$. Si $c_1 + \operatorname{Im}\partial^1 \in \mathrm{N}(\bar{\phi})$ entonces tenemos que $\phi(c_1) \in \operatorname{Im}\partial^2$, digamos $\phi(c_1) = \partial^2(c_2')$, con $c_2' \in Z'^2$. Sea $c_3' = \psi'(c_2')$. Es claro que $c_3' \in \mathrm{N}(\partial^3)$ y por construcción $\delta^*(c_3') = c_1 + \operatorname{Im}\partial^1$, luego concluimos que $c_1 + \operatorname{Im}\partial_1 \in \operatorname{Im}\delta^*$. ■

He aquí el resultado que queríamos probar:

Teorema 11.13 *Si* $0 \longrightarrow \mathcal{A} \xrightarrow{\phi} \mathcal{B} \xrightarrow{\psi} \mathcal{C} \longrightarrow 0$ *es una sucesión exacta de homomorfismos de complejos de grado* 0 *entonces existe un homomorfismo de módulos* $\delta^* : H(\mathcal{C}) \longrightarrow H(\mathcal{A})$ *de grado* 1 *tal que la sucesión siguiente es exacta:*

$$\cdots \longrightarrow H^k(\mathcal{A}) \xrightarrow{\overline{\phi^k}} H^k(\mathcal{B}) \xrightarrow{\overline{\psi^k}} H^k(\mathcal{C}) \xrightarrow{(\delta^*)^k} H^{k+1}(\mathcal{A}) \xrightarrow{\overline{\phi^{k+1}}} H^{k+1}(\mathcal{B}) \longrightarrow \cdots$$

Equivalentemente, tenemos el triángulo exacto

$$\begin{array}{ccc} H(\mathcal{C}) & \xrightarrow{\delta^*} & H(\mathcal{A}) \\ \bar{\psi} \nwarrow & & \swarrow \bar{\phi} \\ & H(\mathcal{B}) & \end{array}$$

Demostración: Tenemos las sucesiones exactas

$$0 \longrightarrow C^k(\mathcal{A}) \xrightarrow{\phi^k} C^n(\mathcal{B}) \xrightarrow{\psi^k} C^k(\mathcal{C}) \longrightarrow 0,$$

para todo $k \in \mathbb{Z}$.

Basta comprobar que el diagrama siguiente se encuentra en las hipótesis del teorema anterior.

$$\begin{array}{ccccccc} & C^k(\mathcal{A})/F^k(\mathcal{A}) & \xrightarrow{\phi^k} & C^k(\mathcal{B})/F^k(\mathcal{B}) & \xrightarrow{\psi^k} & C^k(\mathcal{C})/F^k(\mathcal{C}) \longrightarrow 0 \\ & \partial^k \big\downarrow & & \partial^k \big\downarrow & & \partial^k \big\downarrow \\ 0 \longrightarrow & Z^{k+1}(\mathcal{A}) & \xrightarrow{\phi^{k+1}} & Z^{k+1}(\mathcal{B}) & \xrightarrow{\psi^{k+1}} & Z^{k+1}(\mathcal{C}) \end{array}$$

(donde Z y F representan los grupos de cociclos y cofronteras de los complejos.)

Ciertamente la fila superior está bien definida, ψ^k es suprayectiva y se cumple $\operatorname{Im}\phi^k \subset \mathrm{N}(\psi^k)$.

Si $\psi^k\big(u + F^k(\mathcal{B})\big) = 0$ entonces $\psi^k(u) \in F^k(\mathcal{C})$, luego $\psi^k(u) = \partial^{k-1}(v)$, para un cierto $v \in C^{k-1}(\mathcal{C})$, que a su vez es de la forma $v = \psi^{k-1}(w)$ con $w \in C^{k-1}(\mathcal{B})$. Así pues, $\psi^k(u) = \partial^{k-1}\big(\psi^{k-1}(w)\big) = \psi^k\big(\partial^{k-1}(w)\big)$, con lo que $u - \partial^{k-1}(w) \in \mathrm{N}(\psi^k)$. Por consiguiente existe un $x \in C^k(\mathcal{A})$ tal que $u - \partial^{k-1}(w) = \phi^k(x)$, luego $u + F^k(\mathcal{B}) = \phi^k\big(x + F^k(\mathcal{A})\big)$.

Esto prueba la exactitud de la fila superior. Es obvio que el diagrama conmuta, que ϕ^{k+1} es inyectiva y que $\operatorname{Im}\phi^{k+1} \subset \mathrm{N}(\psi^{k+1})$.

Supongamos por último que $x \in \mathrm{N}(\psi^{k+1})$. Entonces $x = \phi^{k+1}(y)$ para un $y \in C^{k+1}(\mathcal{A})$ y hay que probar que $y \in Z^{k+1}(\mathcal{A})$. Ahora bien, $\phi^{k+2}\big(\partial^{k+1}(y)\big) = \partial^{k+1}\big(\phi^{k+1}(y)\big) = \partial^{k+1}(x) = 0$ (pues x es un ciclo). Como ϕ^{k+2} es inyectiva resulta que $\partial^{k+1}(y) = 0$, luego y es un ciclo. ■

El homomorfismo δ^* recibe el nombre de *homomorfismo de conexión* de la sucesión exacta dada. Conviene recordar cómo actúa: dado un cociclo de $\mathcal{C}$, tomamos cualquier antiimagen por ψ, calculamos la cofrontera de ésta, calculamos su antiimagen por ϕ y la clase del cociclo resultante es la imagen por δ^* de la clase del cociclo de partida.

Veamos un ejemplo importante de aplicación de este teorema:

Sea S una variedad diferenciable y U_1, U_2 dos abiertos en S de modo que $S = U_1 \cup U_2$, $U_1 \cap U_2 \neq \varnothing$. Claramente, U_1, U_2, $U_1 \cap U_2$ son variedades diferenciables. Consideramos las inclusiones

$$j_1 : U_1 \cap U_2 \longrightarrow U_1, \quad j_2 : U_1 \cap U_2 \longrightarrow U_2, \quad i_1 : U_1 \longrightarrow M, \quad i_2 : U_2 \longrightarrow M.$$

A partir de ellas construimos una sucesión de aplicaciones lineales

$$0 \longrightarrow \Lambda(S) \xrightarrow{\alpha} \Lambda(U_1) \oplus \Lambda(U_2) \xrightarrow{\beta} \Lambda(U_1 \cap U_2) \longrightarrow 0. \tag{11.2}$$

Definimos $\alpha(\omega) = \big(i_1^\sharp(\omega), i_2^\sharp(\omega)\big)$ y $\beta(\omega_1, \omega_2) = j_1^\sharp(\omega_1) - j_2^\sharp(\omega_2)$.

Representaremos las diferenciales de $\Lambda(S)$, $\Lambda(U_1)$, $\Lambda(U_2)$ y $\Lambda(U_1 \cap U_2)$ mediante d, d_1, d_2 y d_{12} respectivamente. Es claro que $\Lambda(U_1) \oplus \Lambda(U_2)$ es un complejo con el operador cofrontera dado por $(d_1 \oplus d_2)(\omega_1, \omega_2) = (d_1\omega_1, d_2\omega_2)$. Las aplicaciones α y β son homomorfismos de complejos (es decir, conmutan con las diferenciales), luego inducen aplicaciones lineales

$$\overline{\alpha} : H(S) \longrightarrow H(U_1) \oplus H(U_2), \quad \overline{\beta} : H(U_1) \oplus H(U_2) \longrightarrow H(U_1 \cap U_2).$$

Veamos que la sucesión (11.2) es exacta, con lo que podremos aplicarle el teorema 11.13. En primer lugar probamos que β es suprayectiva. Fijemos una partición de la unidad p_1, p_2 en S subordinada al cubrimiento U_1, U_2, es decir, $p_1 + p_2 = 1$, $p_1 \prec U_1$, $p_2 \prec U_2$.

Tomemos $\omega \in \Lambda(U_1 \cap U_2)$. La función $i_1^\sharp(p_2)$ está definida en U_1 y se anula en un entorno de cada punto de $U_1 \setminus U_2$, luego la forma $\omega_1 = \big(i_1^\sharp(p_2)\big)\omega$ se puede extender a U_1 haciéndola nula en $U_1 \setminus U_2$. Similarmente tenemos $\omega_2 = \big(i_2^\sharp(p_2)\big)\omega \in \Lambda(U_2)$. Es inmediato comprobar que $\omega = \beta(\omega_1, -\omega_2)$.

La inyectividad de α es obvia: si $\alpha(\omega) = 0$ entonces ω se anula en U_1 y en U_2, luego se anula en S.

Es claro que $\alpha \circ \beta = 0$, luego $\operatorname{Im} \alpha \subset \mathrm{N}\,\beta$. Tomemos ahora $(\omega_1, \omega_2) \in \mathrm{N}\,\beta$. Entonces $\omega_1(p) = \omega_2(p)$ para todo $p \in U_1 \cap U_2$ y consecuentemente podemos definir $\omega \in \Lambda(S)$ que extienda simultáneamente a ω_1 y a ω_2, pero esto equivale a decir que $\alpha(\omega) = (\omega_1, \omega_2)$.

Así pues, el teorema 11.13 nos da la existencia de un homomorfismo δ^* de grado 1 que hace conmutativo el triángulo

$$\begin{array}{ccc} H(U_1 \cap U_2) & \xrightarrow{\delta^*} & H(S) \\ \overline{\beta} \nwarrow & & \swarrow \overline{\alpha} \\ & H(U_1) \oplus H(U_2) & \end{array}$$

En otras palabras, tenemos una sucesión exacta

$$\cdots \longrightarrow H^k(S) \xrightarrow{\overline{\alpha}} H^k(U_1) \oplus H^k(U_2) \xrightarrow{\overline{\beta}} H^k(U_1 \cap U_2) \xrightarrow{\delta^*} H^{k+1}(S) \longrightarrow \cdots$$

Esta sucesión se conoce como la *sucesión de Mayer-Vietoris* de S respecto al cubrimiento (U_1, U_2).

Ejemplo Para $n \geq 1$, sea $S^n = \{x \in \mathbb{R}^{n+1} \mid \|x\| = 1\}$, es decir, la esfera de dimensión n. Vamos a calcular su cohomología.

Fijemos $0 < \epsilon < 1$ y consideremos los abiertos

$$U = \{x \in S^n \mid x_{n+1} > -\epsilon\}, \quad V = \{x \in S^n \mid x_{n+1} < \epsilon\}.$$

Los puntos de S^{n+1} con $x_{n+1} = 0$ forman el ecuador de la esfera, el cual divide a la misma en dos hemisferios, correspondientes a $x_{n+1} \leq 0$ y $x_{n+1} \geq 0$ respectivamente. Los abiertos U y V cubren cada uno un hemisferio extendiéndose un poco más allá del ecuador. Podemos formar la sucesión de Mayer-Vietoris de S^n asociada al cubrimiento (U, V):

$$\cdots \longrightarrow H^k(S^n) \longrightarrow H^k(U) \oplus H^k(V) \longrightarrow H^k(U \cap V) \longrightarrow H^{k+1}(S^n) \longrightarrow \cdots$$

Es fácil ver que U y V son contractibles. Por ejemplo, para contraer U hasta el polo norte basta acercar paulatinamente a 1 la coordenada x_{n+1} de cada punto. Por otro lado, $U \cap V$ se puede contraer hasta el ecuador disminuyendo paulatinamente la coordenada x_{n+1} hasta hacerla nula. También es claro que el ecuador de S^n es difeomorfo a S^{n-1} (entendiendo que S^0 está formada por dos puntos). Teniendo en cuenta estas consideraciones, la sucesión de Mayer-Vietoris se reduce a

$$\cdots \longrightarrow H^k(S^n) \longrightarrow H^k(p) \oplus H^k(p) \longrightarrow H^k(S^{n-1}) \longrightarrow H^{k+1}(S^n) \longrightarrow \cdots$$

donde p representa a una variedad de dimensión 0 (un punto). El grupo $H^k(p)$ es distinto según si $k = 0$ o si $k > 0$. Para $k = 0$ tenemos

$$0 \longrightarrow \mathbb{R} \longrightarrow \mathbb{R}^2 \longrightarrow H^0(S^{n-1}) \longrightarrow H^1(S^n) \longrightarrow 0,$$

y para $k > 0$ tenemos

$$0 \longrightarrow H^k(S^{n-1}) \longrightarrow H^{k+1}(S^n) \longrightarrow 0.$$

A partir de la primera sucesión, un simple cálculo de dimensiones nos da la relación $\dim H^1(S^n) = \dim H^0(S^{n-1}) - 1$, con lo que

$$\dim H^1(S^n) = \begin{cases} 1 & \text{si } n = 1 \\ 0 & \text{si } n > 1 \end{cases}$$

La segunda sucesión nos da

$$\dim H^{k+1}(S^n) = \dim H^k(S^{n-1}), \quad \text{para } k > 1.$$

A partir de aquí, una simple inducción prueba que

$$\dim H^k(S^n) = \begin{cases} 1 & \text{si } k = 0 \text{ o } k = n \\ 0 & \text{si } 1 \leq k \leq n - 1 \end{cases}$$

Éste es el mejor resultado que podíamos obtener, teniendo en cuenta que ya sabíamos que $H^n(S^n) \neq 0$. ■

Ejercicio: Calcular la cohomología de un toro.

Ejercicio: Calcular la cohomología de un círculo abierto con n agujeros. Probar que el grupo H^1 de $\mathbb{R}^2$ con infinitos agujeros tiene dimensión infinita.

Veamos un último ejemplo más sofisticado. Sea $J : S \longrightarrow S$ una involución en una variedad, es decir, un difeomorfismo tal que $J \circ J$ sea la identidad. El caso típico es $J : S^n \longrightarrow S^n$ dado por $J(p) = -p$.

Entonces $J^\sharp : \Lambda(S) \longrightarrow \Lambda(S)$ es un automorfismo con la misma propiedad: $J^\sharp \circ J^\sharp = I$. Podemos descomponer $\Lambda(S) = \Lambda_+(S) \oplus \Lambda_-(S)$, donde

$$\Lambda_+(S) = \{\omega \in \Lambda(S) \mid J^\sharp(\omega) = \omega\}, \quad \Lambda_-(S) = \{\omega \in \Lambda(S) \mid J^\sharp(\omega) = -\omega\}.$$

En efecto, basta tener en cuenta que

$$\omega = \frac{\omega + J^\sharp(\omega)}{2} + \frac{\omega - J^\sharp(\omega)}{2}.$$

Estos dos subespacios son estables para la diferencial, luego podemos verlos como complejos inversos, y es claro entonces que $H(S) = H_+(S) \oplus H_-(S)$, donde

$$H_+(S) = \{\alpha \in H(S) \mid \overline{J}(\alpha) = \alpha\}, \quad H_-(S) = \{\alpha \in H(S) \mid \overline{J}(\alpha) = -\alpha\}.$$

Ejemplo Vamos a calcular $H^k_+(S^n)$ y $H^k_-(S^n)$. Obviamente son todos nulos excepto los correspondientes a $k = 0, n$. En cada caso, uno de los dos será nulo y el otro tendrá dimensión 1. Sólo hemos de decidir cuál es cuál. Para $k = 0$ es obvio: la aplicación antípoda J deja invariantes a las funciones constantes, luego $H^0_+(S^n) = \mathbb{R}$ y $H^0_-(S^n) = 0$.

Para $k = n$ sabemos que el elemento de medida dm es un n-cociclo con integral no nula y por lo tanto no es una cofrontera. Por consiguiente la clase

de cohomología de dm es una base de $H^n(S^n)$. Hemos de ver si está en $H^n_+(S^n)$ o en $H^n_-(S^n)$. Dado $p \in S^n$ y $v_1, \ldots, v_n \in T_p(S^n)$, calculamos

$$J^\sharp(dm)(p)(v_1, \ldots, v_n) = dm\big(J(p)\big)\big(dJ(p)(v_1), \ldots, dJ(p)(v_n)\big).$$

Podemos considerar a J definida en todo $\mathbb{R}^{n+1}$. La aplicación J en S es la restricción de ésta, luego dJ en S es la restricción de dJ en $\mathbb{R}^{n+1}$, pero J es lineal, luego $dJ(p) = J$, para todo $p \in \mathbb{R}^n$, luego en definitiva $dJ(p)(v) = -v$. Por consiguiente $J^\sharp(dm)(p) = (-1)^n dm\big(J(p)\big)$.

Notar que la igualdad anterior tiene sentido porque S^n tiene el mismo espacio tangente en dos puntos antípodas. Sin embargo, es fácil ver que la orientación de $T_p(S^n)$ es la opuesta de la de $T_{J(p)}(S^n)$, por lo que $dm\big(J(p)\big) = -dm(p)$. En resumen obtenemos que $J^\sharp(dm) = (-1)^{n+1}dm$, con lo que

$$H_+(S^n) = \begin{cases} \mathbb{R} & \text{si } n \text{ es impar} \\ 0 & \text{si } n \text{ es par} \end{cases} \qquad H_-(S^n) = \begin{cases} 0 & \text{si } n \text{ es impar} \\ \mathbb{R} & \text{si } n \text{ es par} \end{cases}$$

■

El interés de estos grupos de cohomología se debe a lo siguiente:

Teorema 11.14 *Sea $\pi : S \longrightarrow P$ un difeomorfismo local entre variedades, es decir, π es diferenciable y suprayectiva y todo punto de S tiene un entorno abierto V tal que $\pi[V]$ es abierto en P y la restricción de π a V es un difeomorfismo en su imagen. Sea J una involución en S y supongamos que para todo $p \in P$ se cumple $\pi^{-1}(p) = \{q, J(q)\}$, para cierto $q \in S$. Entonces $H^k(P) = H^k_+(S)$.*

Demostración: Basta probar que $\pi^\sharp : \Lambda(P) \longrightarrow \Lambda_+(S)$ es un isomorfismo. Puesto que $J \circ \pi = \pi$, tenemos que $\pi^\sharp \circ J^\sharp = \pi^\sharp$, luego la imagen de $\pi^\sharp$ (que en principio estaría en $\Lambda(S)$) está en $\Lambda_+(S)$.

Para probar que es inyectiva tomemos $\omega \in \Lambda^k(P)$ no nula y veamos que su imagen es no nula. Tenemos que existe $p \in P$ y $v_1, \ldots, v_k \in T_p(P)$ de modo que $\omega(p)(v_1, \ldots, v_k) \neq 0$. Sea $p = \pi(q)$, con $q \in S$. El hecho de que π sea un difeomorfismo local se traduce en que $d\pi(q)$ es un isomorfismo, con lo que existen vectores $w_1, \ldots, w_k$ tales que $d\pi(q)(w_i) = v_i$. Es claro entonces que

$$\pi^\sharp(\omega)(q)(w_1, \ldots, w_k) = \omega(p)(v_1, \ldots, v_k) \neq 0,$$

luego $\pi^\sharp(\omega) \neq 0$.

Tomemos ahora $\omega \in \Lambda^k_+(S)$ y veamos que tiene una antiimagen. Fijemos un punto $p \in P$. Sea $q \in S$ tal que $\pi(q) = p$. Sea V un entorno de q en el cual π sea un difeomorfismo. Sea $\omega_p = (\pi|_V^{-1})^\sharp(\omega|_{\pi[V]})$, que es una k-forma en $\pi[V]$.

Veamos que ω_p no depende de ninguna de las elecciones que hemos hecho para construirla. Si p' es cualquier punto en $\pi[V]$ y q' es su antiimagen en V, entonces

$$\omega_p(p')(v_1, \ldots, v_k) = \omega(q')(d\pi(q')^{-1}(v_1), \ldots, d\pi(q')^{-1}(v_k)). \tag{11.3}$$

Esta expresión no depende más que de ω salvo por el hecho de que hemos escogido la antiimagen q' de p'. Sólo hay otra alternativa, pues p' no tiene más antiimágenes que q' y $J(q')$. Ahora bien, si en (11.3) sustituimos q' por $J(q')$ el miembro derecho se convierte en $\omega(J(q'))$ actuando sobre los vectores $d\pi(J(q'))^{-1}(v_i)$, pero se cumple que $J \circ \pi = \pi$, y por consiguiente $d\pi(q') = dJ(q') \circ d\pi(J(q'))$, luego $d\pi(J(q'))^{-1}(v_i) = dJ(q')\big(d\pi(q')^{-1}(v_i)\big)$ y, en definitiva, el miembro derecho de (11.3) se transforma en $\omega(J(q'))$ actuando sobre los vectores $dJ(q')\big(d\pi(q')^{-1}(v_i)\big)$, pero esto es lo mismo que

$$J^\sharp(\omega)(q')(d\pi(q')^{-1}(v_1), \ldots, d\pi(q')^{-1}(v_k)),$$

que da el mismo resultado, porque $\omega \in \Lambda_+(S)$.

De este modo, para cada punto $p \in P$ hemos construido una forma ω_p en un entorno que al actuar sobre un punto q' cualquiera da un resultado que sólo depende de ω. Por lo tanto, dos formas ω_p y $\omega_{p'}$ coincidirán en su dominio común, luego la familia de formas que hemos definido determinan una única forma $\omega^* \in \Lambda_p(P)$, que en un entorno de cada punto viene dada por (11.3). Es inmediato que $\pi^\sharp(\omega^*) = \omega$. ■

***Ejemplo** El espacio proyectivo $\mathrm{P}^n(\mathbb{R})$ puede identificarse con el espacio que resulta de identificar cada punto de S^n con su antípoda. Como ya hemos comentado en el caso del plano, para considerarlo como variedad diferenciable necesitaríamos un concepto más abstracto de variedad que no exija que las variedades estén contenidas en $\mathbb{R}^m$. De todos modos, $\mathrm{P}^n(\mathbb{R})$ está localmente contenido en $\mathbb{R}^{n+1}$, en el sentido de que una carta de S^n que no cubra más de un hemisferio puede considerarse una carta de $\mathrm{P}^n(\mathbb{R})$, lo que nos permite definir el espacio tangente de cada punto y, en general, todos los conceptos asociados a las variedades diferenciales. No vamos a entrar en detalles, pero si aceptamos que la proyección $\pi : S^n \longrightarrow \mathrm{P}^n(\mathbb{R})$ que a cada par de puntos antípodas les asigna una misma imagen en el espacio proyectivo es un difeomorfismo local, entonces el teorema anterior nos proporciona la cohomología de los espacios proyectivos, que resulta ser:

$$H^0\big(\mathrm{P}^n(\mathbb{R})\big) = \mathbb{R}, \quad H^k\big(\mathrm{P}^n(\mathbb{R})\big) = 0, \quad 1 \leq k < n,$$

$$H^n\big(\mathrm{P}^n(\mathbb{R})\big) = \begin{cases} 0 & \text{si } n \text{ es par} \\ \mathbb{R} & \text{si } n \text{ es impar} \end{cases}$$

Incidentalmente tenemos una prueba de que los espacios proyectivos de dimensión par no son orientables, pues si lo fueran, al ser compactos, deberían cumplir $H^n \neq 0$. ■

11.4 Aplicaciones al cálculo vectorial

Tras los resultados de las secciones anteriores podemos afirmar que los abiertos en $\mathbb{R}^n$ que cumplen condiciones como $H^1(U) = 0$ son bastante frecuentes. Vamos a describir con un poco más de detalle las consecuencias de que los grupos de cohomología sean triviales.

Nota A la hora de dar aplicaciones, resulta conveniente observar que las hipótesis de diferenciabilidad pueden relajarse considerablemente. En efecto, en las secciones anteriores hemos trabajado únicamente con formas de clase C^∞ para que las cuestiones técnicas sobre diferenciabilidad no ocultaran las ideas fundamentales de la cohomología. No obstante, si S es una variedad de clase C^∞ (no vale la pena relajar esta hipótesis) podemos tomar como $\Lambda(S)$ el álgebra de las formas diferenciales continuas y definir $Z^k(S)$ como el espacio de k-formas de clase C^1 con diferencial nula, $F^k(S)$ como el espacio de las diferenciales de $k+1$-formas de clase C^2 y $H^k(S)$ como el correspondiente espacio cociente. En estas condiciones $\Lambda(S)$ no se ajusta a la definición de complejo, pues la diferencial sólo está definida en un subespacio de cada $\Lambda^k(S)$, el formado por las k-formas de clase C^1, pero si el lector repasa las pruebas anteriores se convencerá de que todos los resultados valen en este contexto.[5] Así pues, cuando $H^k(S) = 0$ podemos asegurar que las k-formas cerradas de clase C^1 son exactas. ■

La aplicación más elemental es la siguiente:

Teorema 11.15 *Sea U un abierto en $\mathbb{R}^n$ tal que $H^1(U) = 0$. Entonces un campo $F : U \longrightarrow \mathbb{R}^n$ de clase C^1 es de la forma $F = \nabla f$ para una cierta función $f : U \longrightarrow \mathbb{R}$ si y sólo si*

$$\frac{\partial F_i}{\partial x_j} = \frac{\partial F_j}{\partial x_i}, \quad \textit{para } 1 \leq i < j \leq n.$$

Demostración: Consideremos la 1-forma $\omega = F_1\,dx_1 + \cdots + F_n\,dx_n$. El campo F es el gradiente de una función f si y sólo si $\omega = df$. Por hipótesis esto equivale a $d\omega = 0$ y, como

$$d\omega = \sum_{i=1}^n \sum_{j\neq i} \frac{\partial F_i}{\partial x_j}\,dx_j \wedge dx_i = \sum_{1\leq i<j\leq n} \left(\frac{\partial F_i}{\partial x_j} - \frac{\partial F_j}{\partial x_i}\right) dx_j \wedge dx_i,$$

la condición $d\omega = 0$ equivale a la del enunciado. ■

Observemos que el teorema anterior afirma esencialmente que la condición necesaria que el teorema de Schwarz impone a los campos de gradientes es también suficiente (en los abiertos considerados). Conviene destacar los casos particulares correspondientes a $n = 2$ y $n = 3$:

Teorema 11.16 *Sea U un abierto en $\mathbb{R}^2$ tal que $H^1(U) = 0$. Entonces un campo $F : U \longrightarrow \mathbb{R}^2$ de clase C^1 es conservativo si y sólo si*

$$\frac{\partial f_1}{\partial y} = \frac{\partial f_2}{\partial x}.$$

[5] Las modificaciones son todas obvias. Por ejemplo, en la definición de homomorfismo de complejos hemos de exigir que las formas de clase C^1 se transformen en formas de clase C^1. Lo mismo sucede en la definición de homotopía. Ahora, si h es una homotopía entre dos homomorfismos ϕ y ψ y ω es una k-forma de clase C^1, entonces $\phi(\omega) - \psi(\omega) = d\big(h(\omega)\big)$. Como el miembro izquierdo es de clase C^1 lo mismo le sucede al derecho, luego $\phi(\omega) - \psi(\omega)$ es una cofrontera. Notar que las aplicaciones $i(e_1)$ e I_a^b conservan el grado de derivabilidad y la fórmula (11.1) vale para formas de clase C^1. Esto nos permite probar igualmente el teorema 11.9, y observaciones similares se aplican a los resultados posteriores.

Teorema 11.17 *Sea U un abierto en $\mathbb{R}^3$ tal que $H^1(U) = 0$. Entonces un campo $F : U \longrightarrow \mathbb{R}^3$ de clase C^1 es conservativo si y sólo si* $\operatorname{rot} F = 0$.

Vimos en el capítulo anterior que si F es un campo de clase C^2, entonces $\operatorname{div}\operatorname{rot} F = 0$. El teorema siguiente nos da un recíproco parcial:

Teorema 11.18 *Sea U un abierto en $\mathbb{R}^3$ tal que $H^2(U) = 0$ y sea $F : U \longrightarrow \mathbb{R}^3$ un campo de clase C^1. Entonces existe un campo $G : U \longrightarrow \mathbb{R}^3$ tal que $F = \operatorname{rot} G$ si y sólo si* $\operatorname{div} F = 0$.

DEMOSTRACIÓN: Sabemos que $d\big(d\Phi(F)\big) = \operatorname{div} F\, dm$, por lo que $\operatorname{div} F = 0$ equivale a que $d\Phi(F) = d\omega$, para una cierta 1-forma $\omega = G\, d\vec{r}$, lo cual a su vez equivale a que $F = \operatorname{rot} G$. ■

De aquí se siguen algunos resultados importantes sobre unicidad de un campo:

Teorema 11.19 *Sea $F : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ un campo de clase C^1 tal que $\operatorname{div} F = \operatorname{rot} F = 0$ y F tienda a 0 en infinito. Entonces $F = 0$.*

DEMOSTRACIÓN: Como $\operatorname{rot} F = 0$ tenemos que $F = \nabla\phi$, para cierta función $\phi : \mathbb{R}^3 \longrightarrow \mathbb{R}$ de clase C^2. Como $\Delta\phi = \operatorname{div}\nabla\phi = \operatorname{div} F = 0$, tenemos que ϕ es harmónica. Es inmediato comprobar que si ϕ es cualquier función de clase C^2 se cumple

$$\frac{\partial}{\partial x_i}(\Delta\phi) = \Delta\left(\frac{\partial\phi}{\partial x_i}\right).$$

De aquí se sigue en nuestro caso que las derivadas parciales de ϕ, es decir, las componentes de F son harmónicas. Ahora bien, vimos en el capítulo anterior que una función harmónica que tienda a 0 en infinito ha de ser nula, es decir, $F = 0$. ■

De aquí se sigue que si dos campos en $\mathbb{R}^3$ se anulan en el infinito y tienen la misma divergencia y el mismo rotacional, entonces son iguales. (En realidad basta con que la diferencia tienda a 0 en el infinito.) Es natural preguntarse ahora si las ecuaciones $\operatorname{div} F = G$, $\operatorname{rot} F = H$ tienen solución para dos campos G y H prefijados. Si imponemos a G y H las condiciones necesarias para que F se anule en el infinito la respuesta es afirmativa. Antes conviene probar otro resultado:

Teorema 11.20 *Sea F un campo de clase C^1 en $\mathbb{R}^3$ tal que $\operatorname{div} F$ tenga soporte compacto. Entonces F puede descomponerse como $F = V + U$, donde $\operatorname{rot} V = 0$ y $\operatorname{div} U = 0$.*

DEMOSTRACIÓN: El campo V ha de ser de la forma $\nabla\phi$, para una cierta función ϕ de clase C^2. Además $\Delta\phi = \operatorname{div} V = \operatorname{div} F$. Según vimos en el capítulo anterior, esta ecuación tiene como única solución el potencial newtoniano:

$$\phi(x) = \frac{1}{4\pi}\int_{\mathbb{R}^3} \frac{\operatorname{div} F}{\|x-y\|}\, dm(y).$$

Definimos, pues, ϕ de esta manera y $V = \nabla\phi$. Por lo tanto definimos $U = F - V$. Entonces

$$\operatorname{div} U = \operatorname{div} F - \operatorname{div} V = \Delta\phi - \operatorname{div} \nabla\phi = 0.$$

■

En las condiciones del teorema anterior tenemos que $V = \nabla\phi$ y $U = \operatorname{rot} A$, para una cierta función $\phi : \mathbb{R}^3 \longrightarrow \mathbb{R}$ y un cierto campo $A : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$. A ϕ se le llama *potencial escalar* de F, mientras que A es el *potencial vectorial* de F. El primero está determinado salvo una constante, mientras que el segundo lo está salvo un gradiente. Podemos determinar completamente A si exigimos que $\operatorname{div} A = 0$.

Conviene definir el *laplaciano vectorial* de un campo $A : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ como

$$\Delta A = \nabla \operatorname{div} A - \operatorname{rot} \operatorname{rot} A.$$

Se comprueba que si $A = (A_1, A_2, A_3)$ entonces $\Delta A = (\Delta A_1, \Delta A_2, \Delta A_3)$.

Volviendo a nuestro caso, el potencial vectorial A de un campo F (determinado por la condición $\operatorname{div} A = 0$) cumple $\Delta A = -\operatorname{rot} \operatorname{rot} A = -\operatorname{rot} U = -\operatorname{rot} F$. Concluimos así que los potenciales ϕ y A de un campo F con divergencia de soporte compacto están determinados por las ecuaciones

$$\Delta\phi = \operatorname{div} F, \quad \Delta A = -\operatorname{rot} F.$$

Si $\operatorname{rot} F$ tiene también soporte compacto entonces la última ecuación vectorial equivale a tres ecuaciones escalares análogas a la primera, y las soluciones son los potenciales newtonianos de las componentes del rotacional. En definitiva tenemos

$$F(x) = -\frac{1}{4\pi}\nabla\int_{\mathbb{R}^3} \frac{\operatorname{div} F}{\|x-y\|}\, dm + \frac{1}{4\pi} \operatorname{rot}\int_{\mathbb{R}^3} \frac{\operatorname{rot} F}{\|x-y\|}\, dm.$$

Teorema 11.21 *Dados dos campos $D : \mathbb{R}^3 \longrightarrow \mathbb{R}$ y $R : \mathbb{R}^3 \longrightarrow \mathbb{R}^3$ de clase C^2 y de soporte compacto de modo que $\operatorname{div} R = 0$, existe un único campo vectorial $F : \mathbb{R}^3 \longrightarrow \mathbb{R}$ tal que $\operatorname{div} F = D$ y $\operatorname{rot} F = R$.*

DEMOSTRACIÓN: Basta tomar $\Delta\phi = D$, $\Delta A = -R$, $F = \nabla\phi + \operatorname{rot} A$. ■

Capítulo XII

Funciones Harmónicas

Las funciones harmónicas las introdujimos en el capítulo X, donde obtuvimos algunas de sus propiedades más importantes a partir de las fórmulas de Green. Recordemos que una función f de clase C^2 en un abierto de $\mathbb{R}^n$ es harmónica si cumple la ecuación $\Delta f = 0$, conocida como *ecuación de Laplace*. Las funciones harmónicas aparecen en muchos problemas físicos. Ya hemos visto su relación con los potenciales newtonianos. Por citar otro ejemplo, si V es la velocidad de un fluido incompresible (con densidad constante) sin fuentes ni sumideros ($\operatorname{div} V = 0$) e irrotacional ($\operatorname{rot} V = 0$, lo que equivale a que no forma remolinos), entonces $V = \nabla\phi$, para una cierta función ϕ tal que $\Delta\phi = \operatorname{div} V = 0$, es decir, el campo de velocidades se deriva de un potencial harmónico.

Es obvio que las funciones harmónicas de una variable son exactamente las de la forma $f(x) = ax + b$. Más en general, todas las aplicaciones afines son harmónicas. Las propiedades básicas de las funciones harmónicas pueden verse como generalización de propiedades obvias de las rectas. Por ejemplo, si conocemos los valores que toma una recta en los extremos de un intervalo conocemos también los valores que toma en el interior del mismo. Igualmente sucede que una función harmónica en un abierto acotado está determinada por los valores que toma en la frontera. Esto lo probamos en el capítulo X, al igual que esta relación más concreta:

Teorema 12.1 (Teorema del valor medio de Gauss) *Si $x_0 \in \mathbb{R}^n$ y una función $f : \overline{B_r(x_0)} \longrightarrow \mathbb{R}$ es continua en $\overline{B_r(x_0)}$ y harmónica en $\partial B_r(x_0)$, entonces*

$$f(x_0) = \frac{1}{\sigma(\partial B_r(x_0))} \int_{\partial B_r(x_0)} f(x)\, d\sigma.$$

Esta propiedad generaliza al hecho obvio de que una recta toma en el centro de un intervalo la media aritmética de los valores que toma en sus extremos.

Paralelamente a estos resultados de unicidad, existen resultados de existencia, es decir, dada una función continua sobre la frontera de un abierto Ω, ¿puede extenderse a una función harmónica en Ω y continua $\overline{\Omega}$? Esta cuestión se conoce

como *problema de Dirichlet* para Ω, y entre otras cosas probaremos que tiene solución positiva en una familia muy amplia de abiertos.

12.1 El problema de Dirichlet sobre una bola

En esta sección resolveremos explícitamente el problema de Dirichlet para una bola, es decir, dada una función continua sobre una esfera, veremos cómo extenderla a una función continua sobre la bola que limita y harmónica en su interior. Primeramente demostramos un resultado básico sobre existencia de funciones harmónicas:

Teorema 12.2 *Las únicas funciones harmónicas en $\mathbb{R}^n$ de la forma $g(\|x\|)$ son las de la forma*

$$f(x) = \begin{cases} \dfrac{A}{\|x\|^{n-2}} + B & \textit{si } n \neq 2 \\ A \log \|x\| + B & \textit{si } n = 2 \end{cases}$$

Demostración: Sea f una función de la forma indicada. La función g es de clase C^2 en su dominio, pues f lo es y $g(r) = f(r, 0, \ldots, 0)$. Por consiguiente

$$\frac{\partial f}{\partial x_i} = \frac{dg}{dr}\frac{x_i}{\|x\|}, \qquad \frac{\partial^2 f}{\partial x_i^2} = \frac{d^2 g}{dr^2}\frac{x_i^2}{\|x\|^2} + \frac{dg}{dr}\left(\frac{1}{\|x\|} - \frac{x_i^2}{\|x\|^3}\right),$$

luego

$$\Delta f = \frac{d^2 g}{dr^2} + \frac{dg}{dr}\frac{n-1}{\|x\|} = 0.$$

Esta ecuación se cumple para todo $x \neq 0$ en el dominio de f, de donde se sigue claramente que

$$\frac{d^2 g}{dr^2} + \frac{dg}{dr}\frac{n-1}{r} = 0$$

para todo $r \neq 0$ en el dominio de g. En el ejemplo de la página 247 vimos que las únicas soluciones de esta ecuación son las de la forma

$$g(r) = \begin{cases} \dfrac{A}{r^{n-2}} + B & \text{si } n \neq 2 \\ A \log r + B & \text{si } n = 2 \end{cases}$$

de donde se sigue que f tiene la forma indicada. ∎

Consideremos la bola abierta de centro 0 y radio r en $\mathbb{R}^n$ y una función continua $f : \partial B_r(0) \longrightarrow \mathbb{R}$. Queremos extenderla a una función continua que sea harmónica en $B_r(0)$ (tomamos centro 0 por simplificar la notación, pero todo vale igualmente para un centro arbitrario). Para ello nos valdremos de la función

$$H(x, y) = \frac{\|y\|^2 - \|x\|^2}{\|x - y\|^n}.$$

Una comprobación rutinaria muestra que, para cada $y \in \mathbb{R}^n$ fijo y todo $x \neq y$ se cumple $\Delta_x H = 0$, donde $\Delta_x H$ representa el laplaciano de la función $x \mapsto H(x, y)$. Definimos

$$u_f(x) = \frac{1}{r\sigma_{n-1}} \int_{\|y\|=r} H(x,y)f(y)\, d\sigma(y), \quad \text{para } \|x\| < r,$$

donde σ_{n-1} es la medida de Lebesgue de $\partial B_1(0)$. Como podemos derivar bajo la integral, es claro que Δu_f es harmónica en $B_r(0)$. Vamos a probar que si $z \in \partial B_r(0)$, entonces existe

$$\lim_{x \to z} u_f(x) = f(z).$$

Esto prueba que si extendemos u_f a la frontera de la bola como $u_f(z) = f(z)$ obtenemos una extensión continua de f, que resuelve el problema de Dirichlet. Consideremos primero el caso en que $f = 1$. Entonces

$$u_1(x) = \frac{1}{r\sigma_{n-1}} \int_{\|y\|=r} \frac{r^2 - \|x\|^2}{\|x-y\|^n}\, d\sigma(y).$$

Sea h el giro de centro 0 que cumple $h(\|x\|e_1) = x$, donde e_1 es el primer vector de la base canónica. Aplicando el teorema de cambio de variable resulta que

$$u_1(x) = \frac{1}{r\sigma_{n-1}} \int_{\|y\|=r} \frac{r^2 - \|h(\|x\|e_1)\|^2}{\|h(\|x\|e_1) - h(y)\|^n}\, d\sigma(y) = u_1(\|x\|e_1),$$

luego si llamamos $g(r) = u_1(re_1)$ hemos probado que $u_1(x) = g(\|x\|)$, luego u_1 tiene la forma indicada en el teorema anterior. Ahora bien, u_1 está definida en 0, luego la única posibilidad es que u_1 sea constante. Es fácil ver que $u_1(0) = 1$, luego $u_1 = 1$. Volvamos ahora al caso general. Fijemos un punto tal que $\|z\| = r$.

Como f es continua en z, dado $\epsilon > 0$ existe un $\delta > 0$ tal que si $\|y - z\| < \delta$ entonces $|f(y) - f(z)| < \epsilon/2$. Tomemos $\|x\| < r$ tal que $\|x - z\| \leq \delta/2$. Entonces

$$u_f(x) - f(z) = u_f(x) - f(z)u_1(x) = \frac{1}{r\sigma_{n-1}} \int_{\|y\|=r} \frac{r^2 - \|x\|^2}{\|x-y\|^n} (f(y) - f(z))\, d\sigma(y).$$

Descomponemos en dos partes la integral. La primera sobre el conjunto de los puntos que cumplen $\|y - z\| \geq \delta$ y la segunda sobre los que cumplen $\|y - z\| < \delta$.

En el primer caso tenemos $\delta \leq \|y - z\| \leq \|x - y\| + \|y - z\| \leq \|x - y\| + \delta/2$, luego $\|x - y\| \geq \delta/2$. Así pues, si M es una cota de f,

$$\begin{aligned} |u_f(x) - f(z)| &\leq \frac{2M}{r\sigma_{n-1}}(r^2 - \|x\|^2)\left(\frac{2}{\delta}\right)^n + \frac{\epsilon}{2r\sigma_{n-1}} \int_{\|y\|=r} \frac{r^2 - \|x\|^2}{\|x-y\|^n}\, d\sigma(y) \\ &\leq \frac{2M}{r\sigma_{n-1}}(r^2 - \|x\|^2)\left(\frac{2}{\delta}\right)^n + \frac{\epsilon}{2}. \end{aligned}$$

Si tomamos x suficientemente próximo a z podemos exigir que el primer sumando sea menor que $\epsilon/2$, con lo que obtenemos $|u_f(x) - f(z)| < \epsilon$. Con esto hemos probado el teorema siguiente:

Teorema 12.3 *Sea $f : \partial B_r(x_0) \subset \mathbb{R}^n \longrightarrow \mathbb{R}^n$ una función continua. Entonces existe una única función continua $u_f : \overline{B_r(x_0)} \longrightarrow \mathbb{R}$ que extiende a f y es harmónica en $B_r(x_0)$. En los puntos interiores de la bola u_f viene dada por*

$$u_f(x) = \frac{1}{r\sigma_{n-1}} \int_{\|y-x_0\|=r} H(x-x_0, y)\, f(y)\, d\sigma(y).$$

De aquí se desprenden muchas consecuencias. La unicidad de la extensión nos da inmediatamente la siguiente propiedad de las funciones harmónicas, que generaliza al teorema de Gauss:

Teorema 12.4 *Sea $f : \Omega \longrightarrow \mathbb{R}$ una función harmónica en un abierto de $\mathbb{R}^n$, sea $x_0 \in \Omega$ y $r > 0$ tal que $\overline{B_r(x_0)} \subset \Omega$. Entonces si $\|x\| < r$ se cumple*

$$f(x) = \frac{1}{r\sigma_{n-1}} \int_{\|y-x_0\|=r} H(x-x_0, y-x_0)\, f(y)\, d\sigma(y).$$

El teorema 7.23 puede aplicarse indefinidamente a esta integral, lo que prueba que las funciones harmónicas son siempre de clase C^∞. Las derivadas parciales conmutan obviamente con el laplaciano, luego las derivadas parciales de cualquier orden de una función harmónica son funciones harmónicas. La fórmula anterior nos da una información más precisa sobre las derivadas de una función harmónica, de la que sacaremos consecuencias importantes:

Teorema 12.5 *Sea ω un abierto en $\mathbb{R}^n$, sea $f : \overline{\Omega} \longrightarrow \mathbb{R}$ una función continua en $\overline{\Omega}$ y harmónica en Ω. Para cada punto $x \in \Omega$ sea $d(x) = d(x, \partial\Omega)$. Entonces*

$$\left|\frac{\partial f}{\partial x_i}(x_0)\right| \leq \frac{n}{d(x_0)} \sup_{y\in\partial\Omega} |f(y)|.$$

Demostración: Tomemos $0 < r < d(x_0)$ y apliquemos el teorema anterior en la bola $B_r(x_0)$. Derivando resulta

$$\frac{\partial f}{\partial x_i} = \frac{1}{r\sigma_{n-1}} \int_{\|y-x_0\|=r} \left(\frac{-2(x_i - x_{0i})}{\|x-y\|^n} - n\frac{r^2 - \|x-x_0\|^2}{\|x-y\|^{n+2}}(x_i - y_i) \right) f(y)\, d\sigma,$$

luego

$$\begin{aligned}\left|\frac{\partial f}{\partial x_i}(x_0)\right| &\leq \frac{n}{\sigma_{n-1} r^{n+1}} \int_{\|y-x_0\|=r} |x_{0i} - y_i|\, |f(y)|\, d\sigma \\ &\leq \frac{n}{r} \sup_{\|y-x_0\|=r} |f(y)| \leq \frac{n}{r} \sup_{y\in\partial\Omega} |f(y)|.\end{aligned}$$

La última desigualdad se basa en que f no puede tomar un valor máximo o mínimo en Ω, sino que los valores máximos y mínimos los toma en $\partial\Omega$, como se deduce fácilmente del teorema del valor medio de Gauss. Si $\overline{\Omega}$ no es compacto el supremo puede ser infinito. Puesto que la desigualdad vale para todo $r < d(x_0)$, también se cumple para $d(x_0)$. ∎

Como aplicación probamos lo siguiente:

Teorema 12.6 *Sea Ω un abierto en $\mathbb{R}^n$ y $\{f_n\}_{n=0}^{\infty}$ una sucesión de funciones harmónicas en Ω que converge uniformemente a una función f. Entonces f es harmónica en Ω. Además la sucesión $\{D_i f_n\}$ converge uniformemente a $D_i f$.*

DEMOSTRACIÓN: Tomemos un punto $x \in \Omega$ y apliquemos el teorema anterior a una bola cerrada de centro x contenida en Ω. Puesto que $\{f_n\}_{n=0}^{\infty}$ converge uniformemente en la frontera, es claro que la sucesión $\{D_i f_n\}_{n=0}^{\infty}$ es uniformemente de Cauchy en la bola cerrada, luego converge uniformemente a una función g, que por el teorema 3.28 es $D_i f$. Así pues, f es de clase C^1 en Ω. Como las funciones $\{D_i f_n\}$ también son harmónicas en la bola abierta, el mismo razonamiento prueba que $D_i f$ es de clase C^2. Concretamente, las segundas parciales de f en la bola son el límite uniforme de las segundas parciales de las f_n, luego Δf es el límite de Δf_n, luego $\Delta f = 0$. ■

Otra aplicación importante del teorema 12.5 es el hecho de que una función harmónica en todo $\mathbb{R}^n$ no puede estar acotada:

Teorema 12.7 (Teorema de Liouville) *Si $f : \mathbb{R}^n \longrightarrow \mathbb{R}$ es harmónica y acotada entonces es constante.*

DEMOSTRACIÓN: Aplicamos 12.5 en la bola de centro x y radio r, con lo que obtenemos

$$\left|\frac{\partial f}{\partial x_i}\right| \leq \frac{nM}{r},$$

donde M es una cota de f. Como r es arbitrario concluimos que $\nabla f = 0$, luego f es constante. ■

12.2 Funciones holomorfas

Las funciones harmónicas están estrechamente relacionadas con las funciones holomorfas, que son las funciones de variable compleja análogas a las funciones derivables en $\mathbb{R}$. La definición que enlaza mejor con el cálculo diferencial que estamos estudiando es la siguiente:

Definición 12.8 Sea $\Omega \subset \mathbb{C}^n$ un abierto. Una función $f : \Omega \longrightarrow \mathbb{C}^m$ es *holomorfa* si considerada como aplicación $f : \Omega \subset \mathbb{R}^{2n} \longrightarrow \mathbb{R}^{2m}$ es de clase C^1 y, para cada $z \in \Omega$, la diferencial $df(z) : \mathbb{C}^n \longrightarrow \mathbb{C}^m$ es $\mathbb{C}$-lineal.

A partir de esta definición es obvio que la composición de funciones holomorfas es de nuevo una función holomorfa. También es fácil ver que una función $f : \Omega \longrightarrow \mathbb{C}^m$ es holomorfa si y sólo si lo son sus funciones coordenadas $f_i : \Omega \longrightarrow \mathbb{C}$. Por ello nos limitaremos a estudiar funciones con valores en $\mathbb{C}$.

Veamos qué ha de cumplir una aplicación $\mathbb{R}$-lineal $g : \mathbb{C}^n \longrightarrow \mathbb{C}$ para que sea $\mathbb{C}$-lineal. Si es $\mathbb{C}$-lineal entonces $g(z_1, \ldots, z_n) = c_1 z_1 + \cdots + c_n z_n$, para ciertos números complejos $c_i = a_i + ib_i$. Si llamamos $z_i = x_i + iy_i$ entonces

$$g(x_1, y_1 \ldots, x_n, y_n) = (a_1x_1 - b_1y_1 + \cdots + a_nx_n - b_ny_n, a_1y_1 + b_1x_1 + \cdots + a_ny_n + b_nx_n),$$

luego la matriz de g es

$$\begin{pmatrix} a_1 & b_1 \\ -b_1 & a_1 \\ \vdots & \vdots \\ a_n & b_n \\ -b_n & a_n \end{pmatrix}$$

Recíprocamente, si la matriz de g es de esta forma —para números reales cualesquiera a_i y b_i— entonces g es $\mathbb{C}$-lineal.

Si aplicamos esto al caso en que $g = df(z)$ obtenemos el teorema siguiente:

Teorema 12.9 *Sea $\Omega \subset \mathbb{R}^n$ un conjunto abierto y $f : \Omega \longrightarrow \mathbb{C}$ una función de clase C^1. Entonces f es holomorfa en Ω si y sólo si satisface las* ecuaciones de Cauchy-Riemann*:*

$$\frac{\partial \operatorname{Re} f}{\partial x_i} = \frac{\partial Imf}{\partial y_i}, \quad \frac{\partial \operatorname{Re} f}{\partial y_i} = -\frac{\partial Imf}{\partial x_i}.$$

En tal caso

$$df = \frac{\partial f}{\partial z_1}\, dz_1 + \cdots + \frac{\partial f}{\partial z_n}\, dz_n,$$

donde usamos la notación

$$\frac{\partial f}{\partial z_i} = \frac{\partial \operatorname{Re} f}{\partial x_i} + i\,\frac{\partial \operatorname{Im} f}{\partial x_i}, \quad dz_i = dx_i + i dy_i.$$

Las igualdades entre diferenciales han de entenderse como elementos del espacio $\Lambda^1(\Omega)$ de todas las aplicaciones de Ω en el espacio de aplicaciones $\mathbb{R}$-lineales de $\mathbb{C}^n$ en $\mathbb{C}$ (con las operaciones definidas de forma obvia). Las llamaremos *1-formas complejas.* Para el caso de funciones de una variable es costumbre escribir

$$f'(z) = \frac{df}{dz}(z) \quad \text{en lugar de} \quad \frac{\partial f}{\partial z}(z).$$

Notemos que la derivada parcial respecto de z_i de una función holomorfa f en un punto $(z_1, \ldots, z_n)$ es la derivada de la función $z \mapsto f(z_1, \ldots, z_{i-1}, z, z_{i+1}, \ldots, z_n)$.

Ejemplo Recordemos que la función exponencial compleja viene dada por

$$e^z = e^{x+iy} = e^x(\cos y + i \operatorname{sen} y).$$

Ciertamente se trata de una función de clase C^1 en $\mathbb{C}$ y

$$\frac{\partial \operatorname{Re} e^z}{\partial x} = e^x \cos y, \qquad \frac{\partial \operatorname{Re} e^z}{\partial y} = -e^x \operatorname{sen} y,$$

$$\frac{\partial \operatorname{Im} e^z}{\partial x} = e^x \operatorname{sen} y, \qquad \frac{\partial \operatorname{Im} e^z}{\partial y} = e^x \cos y.$$

Vemos que e^z cumple las ecuaciones de Cauchy-Riemann y por consiguiente es una función holomorfa en $\mathbb{C}$. Su derivada es

$$\frac{de^z}{dz} = \frac{\partial \operatorname{Re} e^z}{\partial x} + i\frac{\partial \operatorname{Im} e^z}{\partial x} = e^x \cos y + i\, e^x \operatorname{sen} y = e^z.$$

■

En general, si f es una función holomorfa en un abierto de $\mathbb{C}$ que extiende a una función real $g : I \longrightarrow \mathbb{R}$, para un cierto intervalo I, es decir, si $\operatorname{Re} f|_I = g$ e $\operatorname{Im} g|_I = 0$, para todo $x \in I$ tenemos que

$$f'(x) = \frac{\partial \operatorname{Re} f}{\partial x}(x) + i\frac{\partial \operatorname{Im} f}{\partial x}(x) = g'(x),$$

es decir, si una función holomorfa f extiende a una función real g, entonces la derivada compleja de f extiende a la derivada real de g.

Llamaremos $\mathcal{H}(\Omega)$ al conjunto de todas las funciones holomorfas en el abierto $\Omega \subset \mathbb{C}^n$. Es claro que la suma de funciones holomorfas es holomorfa, así como el producto de un número complejo por una función holomorfa, más aún, se cumple la relación $d(\alpha_1 f + \alpha_2 g) = \alpha_1 df + \alpha_2 dg$, con lo que $\mathcal{H}(\Omega)$ tiene estructura de $\mathbb{C}$-espacio vectorial. En el caso de una variable tenemos la regla de derivación $(\alpha_1 f + \alpha_2 g)' = \alpha_1 f' + \alpha_2 g'$.

El producto de funciones holomorfas también es una función holomorfa. Para probarlo consideramos la aplicación $f : \mathbb{C}^2 \longrightarrow \mathbb{C}$ dada por $f(z_1, z_2) = z_1 z_2 = x_1 x_2 - y_1 y_2 + i(x_1 y_2 + x_2 y_1)$. Es claro que

$$\frac{\partial \operatorname{Re} f}{\partial x_1} = x_2 = \frac{\partial \operatorname{Im} f}{\partial y_1}, \quad \frac{\partial \operatorname{Re} f}{\partial y_1} = -y_2 = -\frac{\partial \operatorname{Im} f}{\partial x_1},$$
$$\frac{\partial \operatorname{Re} f}{\partial x_2} = x_1 = \frac{\partial \operatorname{Im} f}{\partial y_2}, \quad \frac{\partial \operatorname{Re} f}{\partial y_2} = -y_1 = -\frac{\partial \operatorname{Im} f}{\partial x_2},$$

luego $z_1 z_2$ es holomorfa y $d(z_1 z_2) = z_2\, dz_1 + z_1\, dz_2$. Usando la regla de la cadena deducimos que si $f,\, g \in \mathcal{H}(\Omega)$ entonces $fg \in \mathcal{H}(\Omega)$ y $d(fg) = f\, dg + g\, df$. Esto implica que $\mathcal{H}(\Omega)$ tiene estructura de álgebra. En el caso de funciones de una variable tenemos la regla usual de derivación de productos. Ahora es evidente que el anillo de polinomios $\mathbb{C}[z_1, \ldots, z_n]$ está contenido en $\mathcal{H}(\mathbb{C}^n)$.

Ejercicio: Probar que la función $1/z$ es holomorfa en $\mathbb{C} \setminus \{0\}$ y $d(1/z) = (-1/z^2)\, dz$. Concluir que si $f : \Omega \subset \mathbb{C}^n \longrightarrow \mathbb{C}$ es una función holomorfa que no se anula, entonces $1/f(z)$ es también holomorfa.

Ejercicio: Probar que las funciones $\operatorname{sen} z$ y $\cos z$ son holomorfas, así como que $(\operatorname{sen} z)' = \cos z$, $(\cos z)' = -\operatorname{sen} z$.

Sea $\Omega \subset \mathbb{C}$ un conjunto abierto. Toda 1-forma compleja en Ω es de la forma $\omega = \operatorname{Re}\omega + i \operatorname{Im}\omega$, donde $\operatorname{Re}\omega$ e $\operatorname{Im}\omega$ son dos 1-formas reales en Ω. Si ω es de clase C^1 (es decir, si lo son sus partes real e imaginaria) definimos la 2-forma compleja $d\omega = d\operatorname{Re}\omega + i\, d\operatorname{Im}\omega$. Si $f : \Omega \longrightarrow \mathbb{C}$ es una función holomorfa podemos considerar la 1-forma

$$f(z)\, dz = \big(\operatorname{Re} f(z)\, dx - \operatorname{Im} f(z)\, dy\big) + i\big(\operatorname{Re} f(z)\, dy + \operatorname{Im} f(z)\, dx\big).$$

Las ecuaciones de Cauchy-Riemann implican que $d(f(z)\, dz) = 0$. Por consiguiente, si $H^1(\Omega) = 0$ podemos concluir que existe una 0-forma compleja $g = \operatorname{Re} g + i \operatorname{Im} g$ de clase C^1 en Ω de modo que

$$\operatorname{Re} f(z)\, dx - \operatorname{Im} f(z)\, dy = \frac{\partial \operatorname{Re} g}{\partial x}\, dx + \frac{\partial \operatorname{Re} g}{\partial y}\, dy,$$

$$\operatorname{Re} f(z)\,dy + \operatorname{Im} f(z)\,dx = \frac{\partial \operatorname{Im} g}{\partial x}\,dx + \frac{\partial \operatorname{Im} g}{\partial y}\,dy.$$

Así pues, g cumple las ecuaciones de Cauchy-Riemann y es, por tanto, una función holomorfa. Además $f(z) = g'(z)$. Más aún, $\Delta \operatorname{Re} g = \Delta \operatorname{Im} g = 0$, es decir, $\operatorname{Re} g$ e $\operatorname{Im} g$ son funciones harmónicas, luego también lo son $\operatorname{Re} f$ e $\operatorname{Im} f$. Si $H^1(\Omega) \neq 0$ las conclusiones siguen siendo válidas porque son locales, es decir, podemos restringir f a una bola de centro un punto arbitrario de Ω y así obtenemos que f es de clase C^∞ en dicha bola, que las funciones $\operatorname{Re} f$ e $\operatorname{Im} f$ son harmónicas y además (una vez asegurada la derivabilidad de f') las ecuaciones de Cauchy-Riemann implican que f' es holomorfa.

Más aún, si $f : \Omega \subset \mathbb{C}^n \longrightarrow \mathbb{C}$ es holomorfa y $z = (z_1, \ldots, z_n) \in \Omega$, podemos aplicar el razonamiento anterior a las funciones que fijan todas las componentes de z menos una, con lo cual obtenemos:

Teorema 12.10 *Si $f : \Omega \longrightarrow \mathbb{C}$ es una función holomorfa en un abierto de $\mathbb{C}^n$, entonces f es de clase C^∞ en Ω, $\operatorname{Re} f$ e $\operatorname{Im} f$ son funciones harmónicas y las derivadas parciales de f son también holomorfas.*

Vemos, pues, que las funciones holomorfas son infinitamente derivables. Más aún, todas las propiedades de las funciones harmónicas se traducen a propiedades análogas de las funciones holomorfas. Por ejemplo, el teorema 12.6 nos da:

Teorema 12.11 (Teorema de Weierstrass) *Si $\{f_n\}_{n=0}^{\infty}$ es una sucesión de funciones holomorfas en un abierto $\Omega \subset \mathbb{C}$ que converge uniformemente en los subconjuntos compactos de Ω a una función $f : \Omega \longrightarrow \mathbb{C}$ entonces f es holomorfa en Ω y $\{f_n'\}$ converge a f' uniformemente en los compactos de Ω.*

Basta observar que la convergencia uniforme de $\{f_n\}$ a f equivale a que $\{\operatorname{Re} f_n\}$ converja uniformemente a $\operatorname{Re} f$ e $\{\operatorname{Im} f_n\}$ converja uniformemente a $\operatorname{Im} f$. La posibilidad de relajar la convergencia uniforme de 12.6 a la convergencia uniforme en compactos es clara, y de este modo el teorema es aplicable a las series de potencias (ver 3.26), de donde resulta que todas las funciones definidas por series de potencias son holomorfas. Así tenemos otra prueba de que las funciones e^z, $\operatorname{sen} z$, $\cos z$ son holomorfas.

Ejemplo El teorema de Liouville vale para funciones holomorfas en $\mathbb{C}$, lo cual tiene una aplicación clásica: el teorema fundamental del álgebra. Si un polinomio no constante $P(z)$ no tuviera raíces complejas entonces $1/P(z)$ sería una función holomorfa y acotada en $\mathbb{C}$, lo cual es imposible. ■

Vamos a dar un último resultado sobre funciones de varias variables complejas. Después nos restringiremos al caso de una variable porque es el de mayor interés y las ideas fundamentales se ven así más claramente.

Teorema 12.12 *Sea Ω un abierto en $\mathbb{C}$ tal que $H^1(\Omega) = 0$. Entonces toda función harmónica en Ω es la parte real de una función holomorfa.*

Demostración: Sea $f : \Omega \longrightarrow \mathbb{R}$ una función harmónica. Hemos de probar que existe una función $g : \Omega \longrightarrow \mathbb{R}$ de modo que $f + ig$ sea holomorfa, es decir, que sea de clase C^1 y cumpla

$$\frac{\partial g}{\partial x} = -\frac{\partial f}{\partial y}, \quad \frac{\partial g}{\partial y} = \frac{\partial f}{\partial x}.$$

La existencia de g se sigue inmediatamente del teorema 11.15. ■

Si dos funciones harmónicas f y g cumplen que $f + ig$ es una función holomorfa se dice que son funciones harmónicas *conjugadas.* Es fácil ver que se trata de una relación simétrica y que dos conjugadas de una misma función se diferencian en una constante. El teorema anterior prueba que toda función harmónica en un abierto en $\mathbb{C}$ de cohomología trivial tiene una función harmónica conjugada.

Vamos a necesitar algunos hechos elementales sobre integración de funciones complejas. Supongamos que $C \subset \mathbb{C}$ es una 1-variedad orientable (de hecho, toda 1-variedad es orientable). Si ω es una 1-forma compleja en C, diremos que es *integrable* si lo son $\operatorname{Re}\omega$ e $\operatorname{Im}\omega$, y en tal caso definimos su integral como

$$\int_C \omega = \int_C \operatorname{Re}\omega + i\int_C \operatorname{Im}\omega.$$

Es fácil ver que la integral así definida es $\mathbb{C}$-lineal. Diremos que una función $f : C \longrightarrow \mathbb{C}$ es *integrable* si lo es la 1-forma $f(z)\,dz$, en cuyo caso definimos la integral de f como la de dicha forma.

Antes hemos comprobado que si f es una función holomorfa en un abierto Ω, entonces $d(f\,dz) = 0$, luego aplicando el teorema de Stokes a la parte real y a la parte imaginaria de la integral obtenemos el teorema siguiente:

Teorema 12.13 *Sea Ω un abierto acotado en $\mathbb{C}$ tal que $\overline{\Omega}$ sea una 2-variedad con frontera. Sea f una función holomorfa definida en un abierto que contenga a $\overline{\Omega}$. Entonces*

$$\int_{\partial\Omega} f(z)\,dz = 0.$$

De aquí se desprenden propiedades muy importantes de las funciones holomorfas. Para obtenerlas necesitamos algunos hechos básicos sobre integrales de funciones complejas. Ante todo, a efectos de cálculo es conveniente transformar las integrales sobre 1-variedades en integrales sobre arcos. Supongamos que $\gamma : [a, b] \longrightarrow C$ es una aplicación de clase C^1 (en el sentido de que se extiende a una aplicación de clase C^1 sobre un intervalo que contiene a $[a, b]$) y de modo que su restricción a $]a, b[$ sea una carta de la 1-variedad $C \subset \mathbb{C}$ que cubra todos sus puntos salvo a lo sumo un número finito de ellos.[1] Entonces es fácil ver que

$$\int_C f(z)\,dz = \int_a^b f(\gamma(t))\gamma'(t)\,dt,$$

[1] Todas las integrales sobre una 1-variedad se pueden reducir en la práctica a integrales sobre variedades en estas condiciones. El caso típico es $\gamma(t) = z_0 + re^{it}$, con $t \in [0, 2\pi]$, que cubre a toda la circunferencia de centro z_0 y radio r menos un punto.

donde la última integral se interpreta como el número complejo que se obtiene integrando por separado la parte real y la parte imaginaria del integrando.

Conviene observar que el miembro derecho tiene sentido aunque γ no sea la carta de ninguna variedad. Basta con que γ sea una aplicación de clase C^1 (no necesariamente inyectiva) y f una aplicación continua sobre la imagen de γ. En estas condiciones representaremos la integral por

$$\int_\gamma f(z)\,dz.$$

Como primer hecho importante notamos que el teorema 7.23 es aplicable para derivar integrales de funciones complejas: Si γ es un arco, $K \subset C$ es su imagen y $f : \Omega \times K \longrightarrow \mathbb{C}$ es una función tal que $f(\ ,\zeta)$ es holomorfa en Ω para cada $\zeta \in K$, entonces la función $F(z) = \int_\gamma f(z,\zeta)\,d\zeta$ es holomorfa en Ω y

$$F'(z) = \int_\gamma \frac{df}{dz}(z,\zeta)\,d\zeta.$$

Basta aplicar 7.23 para comprobar que F tiene derivadas parciales y éstas cumplen las ecuaciones de Cauchy-Riemann.

Si $h : [a,b] \longrightarrow \mathbb{C}$ es una función continua se cumple la desigualdad

$$\left|\int_a^b h(t)\,dt\right| \leq \int_a^b |h(t)|\,dt.$$

En efecto: Si la integral de h es nula la desigualdad es obvia. En otro caso sea

$$\alpha = \frac{\left|\int_a^b h(t)\,dt\right|}{\int_a^b h(t)\,dt} \in \mathbb{C}.$$

Entonces

$$\left|\int_a^b h(t)\,dt\right| = \alpha \int_a^b h(t)\,dt = \int_a^b \alpha h(t)\,dt,$$

pero como se trata de un número real, en realidad ha de ser

$$\left|\int_a^b h(t)\,dt\right| = \int_a^b \operatorname{Re}(\alpha h(t))\,dt \leq \int_a^b |h(t)|\,dt,$$

pues $\operatorname{Re} z \leq |z|$ y $|\alpha| = 1$. ∎

Aplicado a la integral de una función f sobre un arco γ tenemos

$$\left|\int_\gamma f(z)\,dz\right| \leq \int_a^b |f(\gamma(t))|\,|\gamma'(t)|\,dt.$$

Ahora notamos que $|\gamma'(t)|\,dt$ es el elemento de longitud (no orientada) de γ, es decir, la medida de Lebesgue. Si convenimos en representarlo por $|dz|$ la desigualdad anterior se escribe mejor de este modo:

$$\left|\int_\gamma f(z)\,dz\right| \leq \int_\gamma |f(z)|\,|dz|.$$

Si γ es una carta que cubre casi todos los puntos de una 1-variedad C, entonces la última integral es simplemente la integral de f respecto a la medida de Lebesgue en C.

Una consecuencia es que si $\{f_n\}_{n=1}^\infty$ es una sucesión de funciones continuas que convergen uniformemente en el rango de un arco γ a una función f, entonces

$$\int_\gamma f(z)\,dz = \lim_n \int_\gamma f_n(z)\,dz.$$

En efecto, dado $\epsilon > 0$ existe un natural n_0 tal que si $n \geq n_0$ entonces $|f_n(z) - f_{n_0}(z)| < \epsilon/L(\gamma)$, donde $L(\gamma)$ es la longitud de γ. Por lo tanto

$$\left|\int_\gamma f(z)\,dz - \int_\gamma f_n(z)\,dz\right| \leq \int_\gamma |f(z) - f_n(z)|\,|dz| \leq \epsilon.$$

■

Una simple comprobación muestra que la regla de Barrow es válida para integrales complejas, es decir,

$$\int_\gamma f'(z)\,dz = f\big(\gamma(b)\big) - f\big(\gamma(a)\big).$$

En particular, si $\gamma(b) = \gamma(a)$ entonces $f'(z)$ tiene integral nula sobre γ.

Hemos visto que si $H^1(\Omega) = 0$ entonces toda función f holomorfa en Ω tiene una primitiva, luego $\int_\gamma f(z)\,dz = 0$, para todo arco cerrado γ en Ω. Esto se conoce como *teorema de Cauchy*. Por supuesto, la hipótesis cohomológica es esencial, como muestra el ejemplo siguiente:

Ejemplo Vamos a probar que si $z_0 \in \mathbb{C}$, $r > 0$ y $n \in \mathbb{Z}$ entonces

$$\int_{|\zeta - z_0| = r} (\zeta - z_0)^n\,d\zeta = \begin{cases} 0 & \text{si } n \neq -1 \\ 2\pi i & \text{si } n = -1 \end{cases}$$

En efecto, se entiende que la integral se calcula sobre la circunferencia de centro z_0 y radio r con la orientación usual (en sentido antihorario), de modo que una carta positiva es $\zeta = z_0 + re^{it}$, para $t \in [0, 2\pi]$, luego

$$\int_{|\zeta - z_0| = r} \frac{1}{\zeta - z_0}\,d\zeta = \int_0^{2\pi} \frac{ire^{it}}{re^{it}}\,dt = 2\pi i.$$

Por otra parte, si $n \neq -1$ la función $(z - z_0)^n$ tiene primitiva en $\mathbb{C} \setminus \{0\}$, concretamente la función

$$\frac{(z - z_0)^{n+1}}{n+1},$$

luego la regla de Barrow implica que la integral es nula. ■

En particular $1/z$ es un ejemplo de función holomorfa en $\mathbb{C} \setminus \{0\}$ que no tiene primitiva en dicho abierto. El resultado que acabamos de probar tiene consecuencias muy importantes. Para empezar tenemos lo siguiente:

Teorema 12.14 (Fórmula integral de Cauchy) *Sea $f : \Omega \longrightarrow \mathbb{C}$ una función holomorfa en un abierto Ω que contenga una bola cerrada $\overline{B_r(z_0)}$. Entonces, si $|z - z_0| < r$ y n es un número natural se cumple*

$$f^{n)}(z) = \frac{n!}{2\pi i} \int_{|\zeta - z_0| = r} \frac{f(\zeta)}{(\zeta - z_0)^{n+1}} d\zeta.$$

Demostración: Probemos primero el caso $n = 0$ y $z = z_0$. En primer lugar observamos que el teorema 12.13 implica que la integral

$$\int_{|\zeta - z_0| = r} \frac{f(\zeta)}{\zeta - z_0} \, d\zeta$$

no depende de r, pues si tomamos dos radios distintos, las circunferencias que determinan (con orientaciones contrarias) son la frontera del anillo comprendido entre ambas, y el integrando es una función holomorfa en la clausura de dicho anillo.

Sea $\epsilon > 0$ y tomemos r suficientemente pequeño para que si $|\zeta - z_0| = r$ entonces $|f(\zeta) - f(z_0)| < \epsilon$. Teniendo en cuenta el ejemplo anterior,

$$\left| \frac{1}{2\pi i} \int_{|\zeta - z_0| = r} \frac{f(\zeta)}{\zeta - z_0} \, d\zeta - f(z_0) \right| = \left| \frac{1}{2\pi i} \int_{|\zeta - z_0| = r} \frac{f(\zeta) - f(z_0)}{\zeta - z_0} \, d\zeta \right|$$

$$\leq \frac{1}{2\pi} \int_{|\zeta - z_0| = r} \frac{|f(\zeta) - f(z_0)|}{r} \, |d\zeta| \leq \frac{\epsilon}{2\pi r} \int_{|\zeta - z_0| = r} |d\zeta| = \epsilon.$$

Como esto se cumple para todo ϵ, concluimos que se da la igualdad buscada. Consideremos ahora un punto z tal que $|z - z_0| < r$ y tomemos un radio s tal que $\overline{B_s(z)} \subset B_r(z_0)$. Entonces el teorema 12.13 junto con la parte ya probada implica que

$$\frac{1}{2\pi i} \int_{|\zeta - z_0| = r} \frac{f(\zeta)}{\zeta - z} \, d\zeta = \frac{1}{2\pi i} \int_{|\zeta - z| = s} \frac{f(\zeta)}{\zeta - z} \, d\zeta = f(z),$$

pues las dos circunferencias constituyen la frontera del abierto $B_r(z_0) \setminus \overline{B_s(z)}$. La fórmula para n arbitrario se obtiene derivando ésta. ■

Introducimos ahora un nuevo concepto que nos permitirá generalizar notablemente estas fórmulas. Consideremos una serie funcional de la forma

$$\sum_{n=1}^{+\infty} \frac{a_{-n}}{(z-z_0)^n}\ , \tag{12.1}$$

donde $a_{-n},\ z_0 \in \mathbb{C}$. La función $h(z) = 1/(z-z_0)$ es un homeomorfismo de $\mathbb{C} \setminus \{z_0\}$ en $\mathbb{C} \setminus \{0\}$ (su inversa es $h^{-1}(z) = z_0 + 1/z$). Es claro que la serie anterior converge (absolutamente) en un punto $z \neq z_0$ si y sólo si la serie de potencias

$$\sum_{n=1}^{\infty} a_{-n} z^n$$

converge (absolutamente) en $h(z)$ (y la suma es la misma). Puesto que esta serie converge absolutamente en un disco de la forma $B_\epsilon(0)$ ($= \mathbb{C}$ si $\epsilon = \infty$) y diverge fuera de la clausura del mismo (teorema 3.26), concluimos que la serie original converge absolutamente en el abierto $A = h^{-1}[B_\epsilon(0)] = \{z \in \mathbb{C} \mid |z-z_0| > r\}$ (donde $r = 1/\epsilon$) y diverge fuera de la clausura del mismo (supuesto $\epsilon > 0$ y con el convenio $1/\infty = 0$). Más aún, si K es un subconjunto compacto de A, la serie de potencias converge uniformemente en $h[K]$, de donde se sigue fácilmente que la serie original converge uniformemente en K. El teorema de Weierstrass implica que (12.1) define una función holomorfa sobre los números z tales que $|z-z_0| > r$. Consideremos ahora una serie de potencias

$$\sum_{n=0}^{\infty} a_n (z-z_0)^n.$$

Si su radio de convergencia R es mayor que r, entonces podemos considerar la función holomorfa

$$f(z) = \sum_{n=-\infty}^{+\infty} a_n (z-z_0)^n = \sum_{n=1}^{+\infty} \frac{a_{-n}}{(z-z_0)^n} \ \ + \sum_{n=0}^{\infty} a_n (z-z_0)^n,$$

definida sobre el anillo $A(z_0, r, R) = \{z \in \mathbb{C} \mid r < |z-z_0| < R\}$, donde quizá $R = +\infty$. Estas series dobles reciben el nombre de *series de Laurent.* Las series

$$\sum_{n=0}^{\infty} a_n (z-z_0)^n, \quad \sum_{n=1}^{\infty} \frac{a_{-n}}{(z-z_0)^n},$$

reciben el nombre de *parte regular* y *parte singular,* respectivamente, de la serie de Laurent.

Los razonamientos anteriores muestran que si una serie de Laurent converge en un abierto (en el sentido de que lo hacen sus partes regular y singular) entonces converge en un anillo $A(z_0, r, R)$. La convergencia es absoluta y uniforme en los compactos. Además la serie diverge fuera de la clausura del anillo.

Ahora es fácil ver que los coeficientes de una serie de Laurent convergente en un anillo están completamente determinados por la función holomorfa que determina:

Teorema 12.15 *Supongamos que una serie de Laurent*

$$\sum_{n=-\infty}^{+\infty} a_n(z-z_0)^n$$

converge en un anillo $A(z_0, r, R)$ a una función f. Sea $r < \rho < R$. Entonces

$$a_n = \frac{1}{2\pi i}\int_{|\zeta - z_0|=\rho} \frac{f(\zeta)}{(\zeta - z_0)^{n+1}} d\zeta.$$

DEMOSTRACIÓN: La serie de Laurent

$$\frac{f(\zeta)}{(\zeta - z_0)^{m+1}} = \sum_{n=-\infty}^{+\infty} a_n(z-z_0)^{n-m-1}$$

converge obviamente en el mismo anillo que f, y la convergencia es uniforme en la circunferencia de radio ρ. Por lo tanto podemos intercambiar la integral con la suma:

$$\int_{|\zeta - z_0|=\rho} \frac{f(\zeta)}{(\zeta - z_0)^{n+1}} d\zeta = \sum_{n=-\infty}^{+\infty} a_n \int_{|\zeta - z_0|=\rho} (z-z_0)^{n-m-1}\, d\zeta = 2\pi i a_m,$$

pues todos los sumandos son nulos excepto el correspondiente a $n = m$, según hemos visto anteriormente. ■

Lo verdaderamente notable es que toda función holomorfa en un anillo admite un desarrollo en serie de Laurent:

Teorema 12.16 *Sea f una función holomorfa en el anillo $A(z_0, r, R)$, donde $z_0 \in \mathbb{C}$ y $0 \leq r < R \leq +\infty$. Sea $r < \rho < R$ y para cada $n \in \mathbb{Z}$ sea*

$$a_n = \frac{1}{2\pi i}\int_{|\zeta - z_0|=\rho} \frac{f(\zeta)}{(\zeta - z_0)^{n+1}} d\zeta.$$

Entonces

$$f(z) = \sum_{n=-\infty}^{+\infty} a_n(z-z_0)^n, \quad \textit{para } r < |z| < R.$$

DEMOSTRACIÓN: El teorema de 12.13 implica que a_n es independiente de la elección de ρ. Dado $z \in A(z_0, r, R)$ tomamos $r < \rho_1 < |z| < \rho_2 < R$. Ahora aplicamos el teorema 12.13 al anillo limitado por las circunferencias de centro z_0 y radios ρ_1 y ρ_2 menos un disco de centro z y radio ϵ suficientemente pequeño para que su clausura esté contenida en dicho anillo. Obtenemos que

$$\begin{aligned} f(z) &= \frac{1}{2\pi i}\int_{|\zeta - z|=\epsilon} \frac{f(\zeta)}{\zeta - z}\, d\zeta \\ &= \frac{1}{2\pi i}\int_{|\zeta - z_0|=\rho_2} \frac{f(\zeta)}{\zeta - z}\, d\zeta - \frac{1}{2\pi i}\int_{|\zeta - z_0|=\rho_1} \frac{f(\zeta)}{\zeta - z}\, d\zeta \end{aligned}$$

$$= \frac{1}{2\pi i}\int_{|\zeta-z_0|=\rho_2}\frac{f(\zeta)}{(\zeta-z_0)-(z-z_0)}\,d\zeta - \frac{1}{2\pi i}\int_{|\zeta-z_0|=\rho_1}\frac{f(\zeta)}{(\zeta-z_0)-(z-z_0)}\,d\zeta$$

$$= \frac{1}{2\pi i}\int_{|\zeta-z_0|=\rho_2}\frac{f(\zeta)}{\zeta-z_0}\frac{d\zeta}{1-\frac{z-z_0}{\zeta-z_0}} + \frac{1}{2\pi i}\int_{|\zeta-z_0|=\rho_1}\frac{f(\zeta)}{z-z_0}\frac{d\zeta}{1-\frac{\zeta-z_0}{z-z_0}}$$

$$\begin{aligned} &= \frac{1}{2\pi i}\int_{|\zeta-z_0|=\rho_2}\frac{f(\zeta)}{\zeta-z_0}\sum_{n=0}^{\infty}\left(\frac{z-z_0}{\zeta-z_0}\right)^n d\zeta \\ &+ \frac{1}{2\pi i}\int_{|z-z_0|=\rho_1}\frac{f(\zeta)}{z-z_0}\sum_{n=0}^{\infty}\left(\frac{\zeta-z_0}{z-z_0}\right)^n d\zeta. \end{aligned}$$

Para intercambiar las sumas con las integrales hemos de justificar que las series convergen uniformemente en las circunferencias. Esto se sigue del criterio de mayoración de Weierstrass. Por ejemplo, para la primera tenemos

$$\left|\frac{f(\zeta)}{\zeta-z_0}\left(\frac{z-z_0}{\zeta-z_0}\right)^n\right| \leq \frac{M}{\rho_2}\left(\frac{|z-z_0|}{\rho_2}\right)^n$$

y la sucesión de la derecha es una progresión geométrica de razón menor que 1, luego determina una serie convergente. Con la segunda serie se razona igualmente. Así pues,

$$\begin{aligned} f(z) &= \sum_{n=0}^{\infty}\frac{1}{2\pi i}\left(\int_{|\zeta-z_0|=\rho_2}\frac{f(\zeta)}{(\zeta-z_0)^{n+1}}d\zeta\right)(z-z_0)^n \\ &+ \sum_{n=0}^{\infty}\frac{1}{2\pi i}\left(\int_{|\zeta-z_0|=\rho_2}f(\zeta)(\zeta-z_0)^n d\zeta\right)(z-z_0)^{-n-1} \\ &= \sum_{n=-\infty}^{+\infty}a_n(z-z_0)^n. \end{aligned}$$

■

Definición 12.17 Sea $f : \Omega \longrightarrow \mathbb{C}$ una función holomorfa. Diremos que un punto z_0 es una *singularidad aislada* de f si $B_r(z_0)\setminus\{z_0\}\subset\Omega$ para cierto radio $r > 0$.

Es decir, z_0 es una singularidad aislada de f si f está definida alrededor de z_0 (pero tal vez no en z_0). Puesto que $B_r(z_0)\setminus\{z_0\} = A(z_0, 0, r)$, el teorema anterior implica que f admite un desarrollo en serie de Laurent

$$f(z) = \sum_{n=-\infty}^{+\infty}a_n(z-z_0)^n, \quad \text{para } 0 < |z-z_0| < r.$$

Los coeficientes a_n están unívocamente determinados por f, luego podemos definir el *orden* de f en z_0 como

$$o(f, z_0) = \text{ínf}\{n\in\mathbb{Z} \mid a_n \neq 0\},$$

entendiendo que $o(f, z_0) = +\infty$ si $a_n = 0$ para todo n y $o(f, z_0) = -\infty$ si hay coeficientes $a_{-n} = 0$ para n arbitrariamente grande.

Veamos la información que nos da el orden de una singularidad aislada. Si $o(f, z_0) \geq 0$ entonces la serie de Laurent es en realidad una serie de potencias, definida también en z_0, por lo que f se extiende a una función holomorfa en $\Omega \cup \{z_0\}$. Concretamente $f(z_0) = a_0$. Más en general, la fórmula de Cauchy nos da que si $n \geq 0$ entonces

$$a_n = \frac{1}{2\pi i} \int_{|\zeta - z_0| = r} \frac{f(\zeta)}{(\zeta - z_0)^{n+1}} d\zeta = \frac{f^{n)}(z_0)}{n!}.$$

Por consiguiente el desarrollo de Laurent de f es

$$f(z) = \sum_{n=0}^{\infty} \frac{f^{n)}(z_0)}{n!} (z - z_0)^n,$$

es decir, la serie de Laurent de f es precisamente su serie de Taylor. Ahora sabemos que la serie converge a f en todo disco abierto de centro z_0 contenido en Ω. Cuando $o(f, z_0) \geq 0$ se dice que z_0 es una *singularidad evitable* de f.

En general, si $o(f, z_0) = n \in \mathbb{Z}$, podemos extraer un término $(z - z_0)^n$ de la serie de Laurent, de modo que nos queda una serie de potencias con primer coeficiente no nulo, es decir, $f(z) = (z - z_0)^n g(z)$, donde $g(z)$ es una función holomorfa definida en un entorno de z_0 y tal que $g(z_0) \neq 0$. La unicidad de la serie de Laurent implica fácilmente que esta descomposición es única.

Si $o(f, z_0) = n > 0$ se dice que z_0 es un *cero* de orden n de f. Notemos que si f es un polinomio este concepto de orden de un cero coincide con la multiplicidad de una raíz. Si $o(f, z_0) = -n < 0$ se dice que z_0 es un *polo* de orden n de z_0. Es claro que $f(z)$ tiende a infinito cuando z tiende a un polo.

Finalmente, si $o(f, z_0) = -\infty$ se dice que el punto z_0 es una *singularidad esencial* de f. Notemos que f no puede estar acotada en un entorno de una singularidad esencial. En efecto, si lo estuviera, también lo estaría la parte singular de su serie de Laurent, es decir, tendríamos una función

$$g(z) = \sum_{n=1}^{+\infty} \frac{a_{-n}}{(z - z_0)^n}$$

convergente para $0 < |z - z_0|$ y acotada en un entorno de z_0. Entonces la función $g(1/(z - z_0))$ está definida por una serie de potencias que converge en $\mathbb{C}$ y tiene que estar acotada en un entorno de ∞, luego está acotada en todo $\mathbb{C}$. Por el teorema de Liouville ha de ser constante, luego de hecho ha de ser nula, al igual que g, lo que nos da una contradicción.

Tampoco puede ser que una función f tienda a ∞ cuando z tiende a una singularidad esencial z_0, pues entonces $1/f$ tendería a 0, luego z_0 sería una singularidad evitable de $1/f$ (no puede ser un polo ni una singularidad esencial), digamos $1/f(z) = (z - z_0)^n g(z)$, para una cierta función g holomorfa en un

entorno de z_0 que no se anula en z_0. Por consiguiente $f(z) = (z-z_0)^{-n}(1/g(z))$, donde el segundo factor es una función holomorfa en un entorno de z_0 que no se anula. Desarrollándola en serie de Taylor vemos que f tiene un polo en z_0, en contra de lo supuesto.

Como los casos que hemos considerado son todos los posibles y se excluyen mutuamente, hemos probado lo siguiente:

Teorema 12.18 *Sea z_0 una singularidad aislada de una función holomorfa f. Entonces*

a) z_0 es una singularidad evitable de f si y sólo si f está acotada en un entorno de z_0. Además en tal caso existe $\lim\limits_{z\to z_0} f(z) \in \mathbb{C}$.

b) z_0 es un polo de f si y sólo si $\lim\limits_{z\to z_0} f(z) = \infty$.

c) z_0 es una singularidad esencial de f si y sólo si f no tiene límite (finito o infinito) en z_0.

Definición 12.19 Si z_0 es una singularidad aislada de una función holomorfa f se llama *residuo* de f en z_0 al coeficiente a_{-1} de su serie de Laurent en z_0. Lo representaremos por $\operatorname{Res}(f, z_0)$.

El residuo de f es lo único que influye al calcular una integral a lo largo de una circunferencia que rodee a z_0 y a ninguna otra singularidad de f. En efecto, si una función f es holomorfa en $B_R(z_0) \setminus \{z_0\}$ y $0 < r < R$ entonces

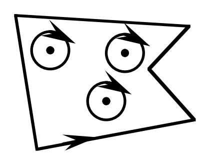
$$\begin{aligned}\int_{|\zeta-z_0|=r} f(\zeta)\,d\zeta &= \int_{|\zeta-z_0|=r} \sum_{n=-\infty}^{+\infty} a_n(z-z_0)^n\,d\zeta \\ &= \sum_{n=-\infty}^{+\infty} a_n \int_{|\zeta-z_0|=r} (z-z_0)^n\,d\zeta \\ &= 2\pi i \operatorname{Res}(f, z_0).\end{aligned}$$

Más en general, tenemos el siguiente resultado, que extiende al teorema de Cauchy.

Teorema 12.20 (Teorema de los residuos) *Sea $\Omega \subset \mathbb{C}$ un abierto acotado tal que $\overline{\Omega}$ sea una variedad con frontera en las condiciones del teorema de Stokes 10.26. Sea f una función holomorfa definida en un abierto que contenga a $\overline{\Omega}$ salvo a lo sumo un número finito de sus puntos, ninguno de ellos en $\partial\Omega$. Entonces*

$$\int_{\partial\Omega} f(\zeta)\,d\zeta = 2\pi i \sum_{z\in\Omega} \operatorname{Res}(f, z).$$

Demostración: Para cada punto $z \in \Omega$ donde no esté definida f tomamos una bola cerrada de centro z y radio r suficientemente pequeño como para que el anillo $A(z, 0, r)$ esté contenido en Ω y no contenga ninguna singularidad de f (no evitable).



Aplicamos el teorema de Stokes a la variedad formada por $\overline{\Omega}$ menos las bolas abiertas. El resultado es que la integral de f en $\partial\Omega$ es igual a la suma de las integrales de f a lo largo de las circunferencias, cuyo valor ya lo hemos calculado y es el que requiere el teorema. ■

El teorema de los residuos tiene innumerables aplicaciones, especialmente al cálculo de integrales y suma de series. Veamos una a modo de ejemplo.

Teorema 12.21 *Consideremos una función racional (cociente de polinomios)*

$$R(x) = \frac{P(x)}{Q(x)} \in \mathbb{R}(x),$$

de modo que $\operatorname{grad} Q(x) \geq \operatorname{grad} P(x) + 2$ *y* $Q(x)$ *no tenga raíces reales. Entonces*

$$\int_{-\infty}^{+\infty} R(x)\, dx = 2\pi i \sum_{\operatorname{Im} z > 0} \operatorname{Res}(R, z).$$

DEMOSTRACIÓN: La función $R(z)$ está definida sobre todo el plano complejo salvo en las raíces de $Q(z)$, que son un número finito. Sea r un número real mayor que el módulo de cualquiera de estas raíces. Consideremos el semicírculo Ω de centro 0 y radio r contenido en el semiplano superior. Está en las condiciones del teorema de Stokes, pues su frontera tiene tan sólo dos puntos singulares ($\pm r$). Al aplicar el teorema de los residuos obtenemos que

$$\int_{\partial\Omega} R(z)\, dz = 2\pi i \sum_{\operatorname{Im} z > 0} \operatorname{Res}(R, z).$$

La frontera de Ω menos sus dos puntos singulares es la unión de dos variedades: la semicircunferencia de carta re^{it}, para $t \in [0, \pi]$, y el intervalo $[-r, r]$, una carta del cual es la identidad. Así pues,

$$\int_{\partial\Omega} R(z)\, dz = \int_{-r}^{r} R(x)\, dx + ir\int_0^{\pi} R(re^{it})e^{it}\, dt. \tag{12.2}$$

Pongamos que

$$R(z) = \frac{a_n z^n + \cdots + a_1 z + a_0}{b_m z^m + \cdots + b_1 z + b_0} = \frac{1}{z^{m-n}} \frac{a_n + \frac{a_{n-1}}{z} + \cdots + \frac{a_1}{z^{n-1}} + \frac{a_0}{z^n}}{b_m + \frac{b_{m-1}}{z} + \cdots + \frac{b_1}{z^{m-1}} + \frac{b_0}{z^m}}.$$

El último término tiende a a_n/b_m cuando z tiende a ∞, luego para valores grandes de $|z|$ se cumple

$$|R(z)| \leq \frac{C}{|z|^{m-n}} \leq \frac{C}{|z|^2}, \quad \text{con } C \in \mathbb{R}.$$

Por consiguiente

$$\left| ir\int_0^{\pi} R(re^{it})e^{it}\, dt \right| \leq \frac{\pi r C}{r^2} = \frac{\pi C}{r},$$

luego el último término de (12.2) tiende a 0 cuando r tiende a $+\infty$. Puesto que la expresión es constante (es la suma de los residuos de $R(z)$), también ha de existir el límite la integral sobre $[-r, r]$. Teniendo en cuenta que $R(x)$ sólo puede cambiar de signo un número finito de veces, el teorema de la convergencia monótona puede aplicarse para justificar que $R(x)$ es integrable en $\mathbb{R}$. Al tomar límites en (12.2) se obtiene la igualdad del enunciado. ■

Ejemplo Vamos a calcular

$$\int_{-\infty}^{+\infty} \frac{dx}{1+x^4}.$$

Si llamamos $\zeta = e^{i\pi/4}$, las raíces del denominador son ζ, ζ^3, ζ^5, ζ^7, de las cuales tienen parte imaginaria positiva

$$\zeta = \frac{\sqrt{2}}{2} + i\,\frac{\sqrt{2}}{2} \quad \text{y} \quad \zeta^3 = -\frac{\sqrt{2}}{2} + i\,\frac{\sqrt{2}}{2}.$$

Tenemos que $R(z) = (z-\zeta)^{-1}g(z)$, donde $g(\zeta) \neq 0$, luego ζ es un polo simple (de orden 1) del integrando R. El argumento que sigue es una técnica general para calcular residuos de polos simples. Ha de ser

$$R(z) = \frac{\operatorname{Res}(R,\zeta)}{z-\zeta} + h(z),$$

para cierta función h holomorfa alrededor de ζ. Por consiguiente

$$(z-\zeta)R(z) = \operatorname{Res}(R,\zeta) + (z-\zeta)h(z)$$

y así el residuo puede calcularse como

$$\operatorname{Res}(R,\zeta) = \lim_{z\to\zeta}(z-\zeta)R(z).$$

En nuestro caso

$$\begin{aligned}
\operatorname{Res}(R,\zeta) &= \lim_{z\to\zeta} \frac{1}{(z-\zeta^3)(z-\zeta^5)(z-\zeta^7)} = \frac{1}{(\zeta-\zeta^3)(\zeta-\zeta^5)(\zeta-\zeta^7)} \\
&= \frac{1}{\zeta^3(1-i)(1+1)(1+i)} = \frac{\zeta^5}{4},
\end{aligned}$$

y del mismo modo llegamos a

$$\operatorname{Res}(R,\zeta^3) = \frac{1}{(\zeta^3-\zeta)(\zeta^3-\zeta^5)(\zeta^3-\zeta^7)} = \frac{\zeta^7}{4}.$$

Por consiguiente

$$\int_{-\infty}^{+\infty} \frac{dx}{1+x^4} = 2\pi i\,\frac{\zeta^5+\zeta^7}{4} = 2\pi i\,\frac{-\sqrt{2}i}{4} = \frac{\pi}{\sqrt{2}}.$$

■

12.3 Funciones subharmónicas

Recordemos que una función real de una variable es convexa si y sólo si su gráfica en un intervalo queda por debajo de la recta que coincide con ella en sus extremos. Si sustituimos "recta" por "función harmónica" y generalizamos a una dimensión arbitrara obtenemos el concepto de función subharmónica:

Definición 12.22 Sea Ω un abierto en $\mathbb{R}^n$. Una función continua $f : \Omega \longrightarrow \mathbb{R}$ es *subharmónica (superharmónica)* si para toda bola cerrada B contenida en Ω se cumple que $f|_B \leq h$ ($f|_B \geq h$), donde h es la función (continua) harmónica en (el interior de) B que coincide con f en ∂B.

Es inmediato que una función es harmónica si y sólo si es subharmónica y superharmónica al mismo tiempo, así como que una función f es subharmónica si y sólo si $-f$ es superharmónica y viceversa. Esto hace que todo teorema sobre funciones subharmónicas se traduzca inmediatamente a otro análogo sobre funciones superharmónicas. Por lo tanto en lo sucesivo trabajaremos únicamente con funciones subharmónicas.

No exigimos que las funciones subharmónicas sean derivables, pero si son al menos de clase C^2 entonces pueden ser caracterizadas en términos de su laplaciano:

Teorema 12.23 *Sea f una función real de clase C^2 en un abierto $\Omega \subset \mathbb{R}^n$. Entonces f es subharmónica si y sólo si $\Delta f \geq 0$.*

Demostración: Supongamos que $\Delta f \geq 0$ y tomemos una bola cerrada B de centro x_0 contenida en Ω. Sea h la función harmónica en B que coincide con f en ∂B. Hemos de probar que $f \leq h$, equivalentemente, que $f|_B - h \leq 0$. Por continuidad y compacidad $f|_B - h$ ha de tomar un valor máximo en la clausura de B. Si éste es positivo lo tomará en un punto $x_1 \in B$ (pues en la frontera f coincide con h). Tomando $c > 0$ suficientemente pequeño, la función

$$\phi(x) = c\|x - x_0\|^2 + f(x) - h(x)$$

cumple $\phi(x_1) > \phi(x)$ para todo $x \in \partial B$. En efecto, si $x \in \partial B$ se cumple $\phi(x) = cr^2$, luego basta tomar $c > 0$ de modo que $cr^2 < f(x_1) - h(x_1)$.

De nuevo por continuidad y compacidad, ϕ tomará su valor máximo en un punto de la clausura de B, pero según lo dicho ha de ser en realidad un punto interior $x_2 \in B$. La función que resulta de fijar todas las variables de ϕ menos la i-ésima tiene un máximo en $(x_2)_i$, luego[2]

$$\frac{\partial^2 \phi}{\partial x_i^2}(x_2) \leq 0.$$

Sumando las derivadas obtenemos que

$$0 \geq \Delta\phi(x_2) = 2nc + \Delta f(x_2) - \Delta h(x_2) = 2nc + \Delta f(x_2),$$

[2]Si una función real f de clase C^2 tiene un máximo en un punto x, entonces $f'(x) = 0$ y $f''(x) \leq 0$, pues si $f''(x) > 0$ tendríamos que f' sería positiva a la derecha de x, con lo que f sería creciente a la derecha de x y $f(x)$ no podría ser un máximo.

luego $\Delta f(x_2) \leq -2nc < 0$, en contra de lo supuesto. Por consiguiente el máximo de $f_B - h$ es menor o igual que 0 y así f es subharmónica.

Recíprocamente, si f es subharmónica en Ω pero $\Delta f < 0$ en algún punto, por continuidad existirá una bola abierta B en la cual $\Delta f < 0$, luego por la parte ya probada $-f$ será subharmónica en B, luego f será subharmónica y superharmónica en B, luego será harmónica y en realidad cumplirá $\Delta f = 0$ en B, con lo que tenemos una contradicción. ■

La propiedad de ser subharmónica es local. El teorema anterior lo prueba para funciones de clase C^2, pero es cierto en general, como se desprende del siguiente resultado:

Teorema 12.24 *Sea $f : \Omega \longrightarrow \mathbb{R}$ una función subharmónica en un abierto de $\mathbb{R}^n$. Si $x_0 \in \Omega$ y $\overline{B_r(x_0)} \subset \Omega$, entonces*

$$f(x_0) \leq \frac{1}{r^{n-1}\sigma_{n-1}} \int_{\|x-x_0\|=r} f(x)\, d\sigma,$$

donde σ_{n-1} es la medida de Lebesgue de la esfera unitaria de dimensión $n-1$.

Recíprocamente, si f es continua y cumple la desigualdad anterior en cada punto x_0 para una sucesión de radios $r_n > 0$ convergente a 0, entonces f es subharmónica en Ω.

Demostración: Sea h la función harmónica en $\overline{B_r(x_0)}$ que coincide con f en la frontera. Entonces

$$f(x_0) \leq h(x_0) = \frac{1}{r^{n-1}\sigma_{n-1}} \int_{\|x-x_0\|=r} f(x)\, d\sigma.$$

Veamos el recíproco. Sea $x_0 \in \Omega$ y sea $R > 0$ tal que $\overline{B_R(x_0)} \subset \Omega$. Sea h la función harmónica en la bola que coincide con f en la frontera. Hemos de probar que $f \leq h$. Sea $g = f - h$ y m su supremo en la bola cerrada. Hemos de ver que es menor o igual que 0. Supongamos, por el contrario, que $m > 0$. Como g es nula en la frontera de la bola, el conjunto $E = \{x \in \overline{B_R(x_0)} \mid g(x) = m\}$ es un subconjunto compacto de $B_R(x_0)$. Sea $x_1 \in E$ tal que $\|x_1 - x_0\|$ sea máximo. De este modo, para todo r suficientemente pequeño, al menos media esfera de centro x_1 y radio r está fuera de E. Tomando como r uno de los valores para los que se cumple la desigualdad del enunciado obtenemos:

$$\begin{aligned} m &= g(x_1) = f(x_1) - h(x_1) \leq \frac{1}{r^{n-1}\sigma_{n-1}} \int_{\|x-x_1\|=r} (f(x) - h(x))\, d\sigma \\ &< \frac{1}{r^{n-1}\sigma_{n-1}} \int_{\|x-x_1\|=r} m\, d\sigma = m, \end{aligned}$$

lo que prueba que $m = 0$, o sea, $f \leq h$ en la bola. Así pues, f es subharmónica. ■

Ejemplo *Si Ω es un abierto en $\mathbb{C}$ y $f : \Omega \longrightarrow \mathbb{C}$ es una función holomorfa, entonces $|f|$ es una función subharmónica.*

En efecto, tomemos una bola $\overline{B_r(z_0)} \subset \Omega$. Entonces la fórmula integral de Cauchy nos da

$$|f|(z_0) = \left| \frac{1}{2\pi i} \int_{|\zeta - z_0| = r} \frac{f(\zeta)}{\zeta - z_0}\, d\zeta \right| \leq \frac{1}{2\pi r} \int_{|\zeta - z_0| = r} |f|(\zeta)\, d\zeta.$$

Basta aplicar el teorema anterior. ■

Veamos otra propiedad importante de las funciones subharmónicas. Para el caso particular del módulo de una función holomorfa recibe el nombre de *Principio del módulo máximo.*

Teorema 12.25 (Principio del máximo) *Sea f una función subharmónica no constante en un abierto $\Omega \subset \mathbb{R}^n$. Entonces*

a) Para todo $x_0 \in \Omega$ se cumple $f(x_0) < \sup\limits_{x \in \Omega} f(x)$.

b) Si Ω está acotado y f es continua en $\overline{\Omega}$, entonces para todo $x_0 \in \Omega$ se cumple

$$f(x_0) < \max_{x \in \partial\Omega} f(x).$$

Demostración: Podemos suponer que Ω es conexo. Sea $m = \sup\limits_{x \in \Omega} f(x)$ (quizá $m = +\infty$). Descomponemos Ω como unión de los conjuntos

$$\Omega_1 = \{x \in \Omega \mid f(x) = m\}, \quad \Omega_2 = \{x \in \Omega \mid f(x) < m\}.$$

La continuidad de f implica que Ω_2 es abierto. Si probamos que Ω_1 también lo es, por conexión uno de los dos será vacío, pero como f no es constante tendrá que serlo Ω_1 y a) quedará demostrado. Para probar que Ω_1 es abierto podemos suponer que es no vacío, lo que implica que m es finito.

Tomemos $x_0 \in \Omega_1$. Al ser f subharmónica, para r suficientemente pequeño se cumple

$$0 \leq \int_{\|x - x_0\| = r} f(x)\, d\sigma - \sigma_{n-1} r^{n-1} f(x_0) = \int_{\|x - x_0\| = r} \big(f(x) - f(x_0)\big)\, d\sigma$$

Como $f(x) - f(x_0) = f(x) - m \leq 0$, la desigualdad anterior es una igualdad, y $f(x) = m$ para todo x tal que $\|x - x_0\| = r$, para todo r suficientemente pequeño, es decir, hay un entorno de x_0 contenido en Ω_1. Esto prueba a). El apartado b) es una consecuencia inmediata. ■

Como aplicación obtenemos la unicidad de la solución del problema de Dirichlet para abiertos con frontera no vacía (no necesariamente acotados):

Teorema 12.26 *Sea Ω un abierto en $\mathbb{R}^n$ distinto de $\varnothing$ y de $\mathbb{R}^n$. Sea $f : \overline{\Omega} \longrightarrow \mathbb{R}$ una función continua harmónica en Ω y tal que $f = 0$ en $\partial\Omega$. Entonces $f = 0$ en Ω.*

DEMOSTRACIÓN: Basta probar que f es constante en Ω, pero si no lo fuera, por el teorema anterior debería ser $f < 0$ en Ω (porque f es subharmónica) y $f > 0$ en Ω (porque $-f$ es subharmónica). ■

Veamos algunas propiedades adicionales que en la sección siguiente nos ayudarán a probar que el problema de Dirichlet tiene solución en una familia muy amplia de abiertos.

Teorema 12.27 *El máximo de dos funciones subharmónicas es una función subharmónica.*

DEMOSTRACIÓN: Sean f_1 y f_2 dos funciones subharmónicas en un abierto $\Omega \subset \mathbb{R}^n$. Sea $f(x) = max\{f_1(x), f_2(x)\}$. Fijada una bola cerrada B contenida en Ω, sean h, h_1 y h_2 las funciones harmónicas en en B que coinciden con f, f_1 y f_2 respectivamente en la frontera. Entonces $h - h_1 \geq 0$ en ∂B, y al ser harmónica la desigualdad vale también en B, es decir, $f_1 \leq h_1 \leq h$, e igualmente $f_2 \leq h_2 \leq h$, luego $f \leq h$. Por consiguiente f es subharmónica. ■

Similarmente se prueba:

Teorema 12.28 *La suma de dos funciones subharmónicas es una función subharmónica.*

Teorema 12.29 *Sea Ω un abierto en $\mathbb{R}^n$ y B una bola cerrada contenida en Ω. Sea f una función subharmónica en Ω y f' la función que coincide con f fuera de B y es harmónica en B. Entonces f' es subharmónica en Ω.*

DEMOSTRACIÓN: Aplicaremos el teorema 12.24. Basta considerar puntos $x_0 \in \partial B$. Notemos que $f \leq f'$ en Ω. Entonces

$$f'(x_0) = f(x_0) = \frac{1}{\sigma_{n-1} r^{n-1}} \int_{\|x-x_0\|=r} f(x)\, d\sigma \leq \frac{1}{\sigma_{n-1} r^{n-1}} \int_{\|x-x_0\|=r} f'(x)\, d\sigma.$$

■

12.4 El problema de Dirichlet

Recordemos que el problema de Dirichlet para un abierto $\Omega \subset \mathbb{R}^n$ consiste en determinar si las funciones continuas en $\partial\Omega$ pueden extenderse a funciones continuas en $\overline{\Omega}$ harmónicas en Ω. El teorema 12.26 prueba que tales extensiones son únicas cuando realmente existen (suponiendo $\partial\Omega \neq \varnothing$), pero hasta ahora sólo hemos probado que el problema tiene solución para las bolas abiertas.

Sea $f : \partial\Omega \longrightarrow \mathbb{R}$ una función continua y acotada. Llamaremos *familia de Perron* de f al conjunto $P(f, \Omega)$ de todas las funciones u continuas en $\overline{\Omega}$, subharmónicas en Ω y tales que $u|_{\partial\Omega} \leq f$. Si M es una cota de f, entonces el principio del máximo prueba que toda función u en estas condiciones cumple $u \leq M$ en Ω. Por consiguiente podemos definir la *función de Perron* de f como

$$P_f(x) = \sup\{u(x) \mid u \in P(f, \Omega)\}, \quad \text{para } x \in \Omega.$$

Ahora observamos que si el problema de Dirichlet tiene solución para f y Ω, entonces la solución ha de ser P_f. En efecto, la solución g está obviamente en $P(f, \Omega)$, luego $g \leq P_f$. Por otra parte, si $u \in P(f, \Omega)$ la función $u - g$ es subharmónica y en $\partial\Omega$ es ≤ 0, luego por el principio del máximo se cumple $u - g \leq 0$ en Ω, o sea, $u \leq g$, luego $P_f \leq g$.

Para probar que P_f es realmente solución del problema de Dirichlet hemos de ver dos cosas: que es harmónica y que tiende a f en $\partial\Omega$. Lo primero podemos probarlo ya:

Teorema 12.30 *Sea Ω un abierto en $\mathbb{R}^n$ tal que $\partial\Omega \neq \varnothing$. Sea $f : \partial\Omega \longrightarrow \mathbb{R}$ una función continua y acotada. Entonces P_f es harmónica en Ω.*

Demostración: Tomemos $x_0 \in \Omega$ y sea $r > 0$ tal que $\overline{B}_r(x_0) \subset \Omega$. Podemos tomar una sucesión de funciones $u_n \in P(f, \Omega)$ tales que $\lim\limits_n u_n(x_0) = P_f(x_0)$. Sustituyendo cada u_n por el máximo de todas las anteriores podemos suponer que u_n es creciente. Más aún, aplicando el teorema 12.29 obtenemos funciones $v_n \in P(f, \Omega)$ que son harmónicas en $B_r(x_0)$. La sucesión v_n también es creciente y, como $u_n(x_0) \leq v_n(x_0) \leq P_f(x_0)$, la sucesión $v_n(x_0)$ también converge a $P_f(x_0)$.

Veamos que v_n converge uniformemente en un entorno de x_0. Esto probará que P_f es harmónica en un entorno de x_0. Sea $0 < r' < r$. Sea $\|y - x_0\| = r$ y $\|x - x_0\| = r'$. Entonces $r - r' \leq \|y - x\|$, luego

$$\frac{r^2 - r'^2}{\|y - x\|^n} \leq \frac{r^2 - r'^2}{(r - r')^n},$$

y el teorema 12.3 implica que

$$v_{n+1}(x) - v_n(x) \leq \frac{r^2 - r'^2}{(r - r')^n} \frac{1}{r\sigma_{n-1}} \int_{\|y - x_0\| = r} (v_{n+1} - v_n)\, d\sigma,$$

luego el teorema del valor medio nos da

$$v_{n+1}(x) - v_n(x) \leq \frac{r^2 - r'^2}{(r - r')^n} r^{n-2} (v_{n+1}(x_0) - v_n(x_0)).$$

Fijado $0 < k < r$, para todo $x \in B_k(x_0)$ se cumple $r' < k$, luego

$$\frac{r^2 - r'^2}{(r - r')^n} r^{n-2} = \frac{r + r'}{(r - r')^{n-1}} r^{n-2} \leq \frac{r + k}{(r - k)^{n-1}} r^{n-2} = M,$$

con lo que $v_{n+1}(x) - v_n(x) \leq M(v_{n+1}(x_0) - v_n(x_0))$. Puesto que

$$v_0(x_0) + \sum_{n=0}^{\infty} (v_{n+1}(x_0) - v_n(x_0)) = P_f(x_0),$$

el criterio de mayoración de Weierstrass implica que la serie

$$v_0(x) + \sum_{n=0}^{\infty} (v_{n+1}(x) - v_n(x))$$

converge uniformemente en $B_k(x_0)$, pero dicha serie no es sino la sucesión v_n.

■

En general no es cierto que P_f coincida con f en $\partial\Omega$. Por ejemplo, tomemos $\Omega = B_1(0) \setminus \{(0,0)\} \subset \mathbb{R}^2$ y sea f la función que vale 0 en $\partial B_1(0)$ y $f(0) = 1$.

Tomemos $u \in P(f,\Omega)$. Por el principio del máximo se cumple que $\|u\| \leq 1$. Dado $0 < \epsilon < 1$, la función $h_\epsilon(x) = (\log\|x\|)/\log\epsilon$ es harmónica en el anillo comprendido entre las circunferencias de centro 0 y radios ϵ y 1 (teorema 12.2). Además h_ϵ vale 0 sobre la circunferencia exterior y 1 sobre la interior. En consecuencia $u \leq h_\epsilon$ en la frontera del anillo. Como $u - h_\epsilon$ es subharmónica, de hecho $u \leq h_\epsilon$ en todo el anillo (por el principio del máximo), es decir,

$$u(x) \leq \frac{\log\|x\|}{\log\epsilon},$$

para $0 < \epsilon \leq \|x\| < 1$. Si fijamos x y hacemos tender ϵ a 0 queda $u(x) \leq 0$ para todo $x \in B_1(0) \setminus \{0\}$ y toda $u \in P(f,\Omega)$. Por consiguiente $P_f = 0$ y no converge a f en 0. No existe ninguna función harmónica en Ω continua en $\overline{B_1(0)}$ que tome el valor 0 para $\|x\| = 1$ y el valor 1 en $x = 0$. ■

Veamos una condición necesaria para que las funciones continuas y acotadas en la frontera de un abierto Ω se extiendan a funciones harmónicas en Ω. Cuando Ω tiene esta propiedad (entendiendo que $\partial\Omega \neq \varnothing$) se dice que es una *región de Dirichlet*. Dado un punto $a \in \partial\Omega$, podemos considerar la función

$$f(x) = \frac{\|x-a\|}{1+\|x-a\|}.$$

Claramente se trata de una función continua y acotada en $\mathbb{R}^n$ (toma valores en $[0,1]$) y que se anula únicamente en a. Si Ω es una región de Dirichlet existe una función continua $h : \overline{\Omega} \longrightarrow \mathbb{R}$ harmónica en Ω y que coincide con f en $\partial\Omega$. En particular $h(a) = 0$ y $h(x) > 0$ para todo $x \in \partial\Omega$, $x \neq a$. Vamos a probar que una condición más débil que ésta es también suficiente para que Ω sea una región de Dirichlet.

Definición 12.31 Sea Ω un abierto en $\mathbb{R}^n$ distinto de $\varnothing$ y de $\mathbb{R}^n$. Para cada $\epsilon > 0$ y $a \in \partial\Omega$ sea $\Omega_\epsilon(a) = \Omega \cap B_\epsilon(a)$. Una *barrera* para Ω en a es una función continua $u : \overline{\Omega_\epsilon(a)} \longrightarrow \mathbb{R}$ subharmónica en $\Omega \cap B_\epsilon(a)$ tal que $u(a) = 0$ y $u(x) < 0$ para todo $x \in \partial\Omega_\epsilon(a)$, $x \neq a$.

Es claro que si Ω es una región de Dirichlet entonces tiene una barrera en cada punto de su frontera (para cualquier $\epsilon > 0$, restringimos a $\overline{\Omega_\epsilon(a)}$ la función $-h$ que hemos construido en el párrafo anterior).

Las regiones de Dirichlet admiten la caracterización siguiente:

Teorema 12.32 *Un abierto $\Omega \subset \mathbb{R}^n$ de frontera no vacía es una región de Dirichlet si y sólo si tiene una barrera en cada punto de su frontera.*

Demostración: Ya hemos visto que la condición es necesaria. Veamos que también es suficiente. Para ello consideremos una función continua y acotada $f : \partial\Omega \longrightarrow \mathbb{R}$. Basta probar que P_f es continua en $\overline{\Omega}$ y que coincide con f en $\partial\Omega$. Fijemos un punto $a \in \partial\Omega$ y sea $v : \overline{\Omega_\epsilon(a)} \longrightarrow \mathbb{R}$ una barrera para Ω en a.

Por el principio del máximo $v < 0$ en $\Omega_\epsilon(a)$. Tomando $0 < r < \epsilon$ tenemos que $v < 0$ en el compacto $\overline{\Omega_\epsilon(a)} \setminus B_r(a)$, luego existe un $\eta > 0$ tal que $v \leq -\eta$ en dicho compacto.

Sea $w : \overline{\Omega_\epsilon(a)} \longrightarrow \mathbb{R}$ el máximo entre $-\eta$ y v. Entonces w es también una barrera para Ω en a con la propiedad adicional de que vale $-\eta$ fuera de $B_r(a)$, luego se puede extender de forma constante a $w : \overline{\Omega} \longrightarrow \mathbb{R}$ y sigue cumpliendo las propiedades de una barrera (salvo que su dominio es mayor). Concretamente, w es continua en $\overline{\Omega}$, subharmónica en Ω, $w(x) < 0$ para todo $x \in \partial\Omega$, $x \neq a$ y $w(a) = 0$.

Sean K, ϵ constantes positivas y consideremos la función

$$u(x) = f(a) - \epsilon + Kw(x).$$

Claramente u es continua en $\overline{\Omega}$ y subharmónica en Ω. Además $u(x) \leq f(a)-\epsilon$ para todo $x \in \partial\Omega$ y $u(a) = f(a) - \epsilon$. Como f es continua en a, existe un $\delta > 0$ tal que $f(x) > f(a) - \epsilon$ para los $x \in \partial\Omega$ que cumplen $\|x - a\| < \delta$. Así pues, $u(x) \leq f(x)$, para estos puntos x. Como w tiene una cota superior negativa en el conjunto de puntos $x \in \partial\Omega$ tales que $\|x - a\| \geq \delta$, eligiendo K adecuadamente podemos exigir que $u(x) \leq f(x)$ para todo $x \in \partial\Omega$. Por consiguiente $u \in P(f, \Omega)$, luego $u \leq P_f$.

Por consiguiente $f(a) - \epsilon = u(a) \leq P_f(a)$, para todo $\epsilon > 0$, lo que prueba que $P_f(a) = f(a)$. No obstante esto no prueba la continuidad de P_f en a. Puesto que $\lim\limits_{x \to a} u(x) = f(a) - \epsilon$, en realidad hemos visto que para puntos $x \in \overline{\Omega}$ suficientemente próximos a a se cumple $f(a) - 2\epsilon \leq P_f(x)$.

Aplicamos este hecho a la función P_{-f}, es decir, sabemos que para puntos $x \in \overline{\Omega}$ suficientemente próximos a a se cumple $-f(a) - 2\epsilon \leq P_{-f}(x)$.

Ahora bien, si $v_1 \in P(f, \Omega)$ y $v_2 \in P(-f, \Omega)$ entonces $v_1 + v_2 \leq 0$ en $\partial\Omega$ y es subharmónica en Ω, luego $v_1 + v_2 \leq 0$ en Ω, luego $v_2 \leq -v_1$ y tomando el supremo $P_{-f} \leq -v_1$, luego $v_1 \leq -P_{-f}$, luego tomando de nuevo el supremo $P_f \leq -P_{-f}$. Cambiando f por $-f$ obtenemos también la otra desigualdad, es decir,

$$f(a) - 2\epsilon \leq P_f(x) \leq -P_{-f}(x) \leq f(a) + 2\epsilon,$$

para puntos $x \in \overline{\Omega}$ suficientemente próximos a a, es decir, existe

$$\lim_{x \to a} P_f(x) = f(a).$$

■

Terminamos observando que la existencia de barreras es una condición que se cumple en una gran clase de abiertos. Por ejemplo, si Ω tiene un hiperplano tangente en un punto $a \in \partial\Omega$, es decir, si todos los puntos de $\overline{\Omega}$ están en un mismo lado del hiperplano excepto el propio a, que está en el mismo, podemos tomar como barrera una aplicación afín que se anule en el hiperplano.

Más en general, si existe una bola cerrada B tal que $B\cap\overline{\Omega} = \{a\}$ entonces Ω tiene una barrera en a. Basta considerar una función de la forma $A/\|x-c\|+B$, donde c es el centro de B.

Capítulo XIII

Aplicaciones al electromagnetismo

La electricidad y el magnetismo son conocidos desde la antigüedad, si bien no han sido comprendidos satisfactoriamente hasta hace relativamente poco tiempo, en la culminación de una investigación que comenzó en el siglo XVIII con el trabajo de Coulomb y terminó con el descubrimiento de la teoría de la relatividad. En la teoría electromagnética clásica (prerrelativista) se aplican muchos de los resultados que hemos obtenido en los últimos capítulos, por lo que constituye un complemento idóneo a los mismos. Toda la teoría puede deducirse a partir de las llamadas ecuaciones de Maxwell, pero en lugar de introducirlas directamente dedicaremos las dos primeras secciones a presentar los conceptos que involucran en un orden más cercano al proceso histórico que condujo hasta ellas.

13.1 Electrostática

Con su balanza de torsión, Coulomb estableció en 1785 la ley fundamental de la electrostática, que lleva su nombre: las fuerzas eléctricas aparecen en relación con una magnitud de la materia llamada *carga eléctrica.* Existen dos clases de carga eléctrica, que arbitrariamente podemos llamar *positiva* y *negativa.* Dos partículas con carga eléctrica experimentan una fuerza proporcional al producto de sus cargas e inversamente proporcional al cuadrado de la distancia que las separa. La fuerza es atractiva si las cargas tienen signos opuestos, y es repulsiva si los signos son iguales.

Observamos que la ley de Coulomb es muy similar a la ley de la gravitación de Newton, salvo por el hecho de que la fuerza gravitatoria es siempre atractiva. Existe otra diferencia importante: la masa de un cuerpo puede definirse sin hacer referencia a la gravitación, como la proporción entre la fuerza que se le aplica y la aceleración que experimenta. El hecho de que esta magnitud de *masa inerte* coincida con la *masa gravitatoria* que aparece en la ley de Newton no es

algo obvio en absoluto, sino una propiedad importante de la gravitación. En particular, la constante gravitatoria G que aparece en la ley de Newton es una constante física que ha de ser medida para ajustar la fórmula. Por el contrario, la carga eléctrica sólo se manifiesta en la medida en que los cuerpos experimentan fuerzas eléctricas, por lo que no tenemos predefinida una unidad de carga. Así pues, podríamos definir la unidad de carga eléctrica como la que hace que dos cargas unitarias separadas por una distancia de 1 metro experimenten una fuerza de 1 Newton, y en tal caso la constante que requiere la ley de Coulomb tomaría el valor 1. Por razones de tradición, la unidad de carga eléctrica, llamada *culombio,* se define como la que hace que dos cargas unitarias separadas por una distancia de un metro experimenten una fuerza de $9 \cdot 10^9$ Newtons.

Además es costumbre llamar

$$\epsilon_0 = \frac{1}{36\pi 10^9},$$

con lo que la constante que requiere la ley de Coulomb[1] tiene la forma $1/4\pi\epsilon_0$.

Ahora ya podemos enunciar matemáticamente la ley de Coulomb en su forma habitual: Si dos partículas puntuales de cargas q y Q están situadas en las posiciones x e y respectivamente, entonces, la fuerza eléctrica que Q ejerce sobre q es

$$F = \frac{1}{4\pi\epsilon_0}\frac{Qq}{\|x-y\|^3}(x-y).$$

Si comparamos la fórmula anterior con la ley de Newton veremos que falta un signo negativo. La razón es que dos masas del mismo signo (o sea, dos masas cualesquiera) se atraen, mientras que las cargas del mismo signo se repelen. Esto hace que todas las fórmulas subsiguientes difieran en un signo de sus análogas gravitatorias.

En este punto conviene hacer algunas observaciones sobre la forma en que las cargas eléctricas aparecen en la naturaleza. La materia ordinaria está compuesta por átomos, cada uno de los cuales consta de un núcleo cargado positivamente y una corteza de electrones cargados negativamente (los signos se estipulan por convenio), de modo que la carga total del átomo es nula. Los electrones tienen cierta facilidad de pasar de unos átomos a otros, de modo que en ocasiones podemos encontrarnos con átomos con un exceso o defecto de electrones *(iones),* pero incluso entonces los iones se encuentran mezclados (aleatoriamente en el caso de sales disueltas o según patrones geométricos en el caso de las estructuras cristalinas), de modo que en cualquier volumen macroscópico de materia la carga total es nula.

En principio, la ley de Coulomb, tal y como la hemos establecido, es válida para dos cargas puntuales separadas por un espacio vacío. Si las tenemos separadas por un medio material la presencia de cargas eléctricas hace que la ley de Coulomb no sea aplicable tal cual. Además es imposible tener en cuenta

[1] Actualmente el culombio se define de otra forma más fácil de medir, con lo que este valor resulta ser aproximado.

cada una de las innumerables cargas eléctricas del medio para estudiar sus efectos. Afortunadamente, cuando las cargas del medio están distribuidas de forma homogénea, sin ninguna estructura destacada, como ocurre en un fluido (aire, agua, etc.) o el seno de un metal —y éstos son los casos de mayor interés— su efecto consiste únicamente en una atenuación de las fuerzas eléctricas, de modo que la ley de Coulomb sigue siendo válida sin más que sustituir la constante ϵ_0 por otra constante ϵ dependiente del medio. Concretamente se suele escribir $\epsilon = \epsilon_r\epsilon_0$, donde ϵ recibe el nombre de *permisividad* del medio, ϵ_r es la *permisividad relativa* del mismo y, naturalmente, ϵ_0 es la permisividad del vacío.

En el seno de estructuras materiales más complejas, como las sustancias cristalinas, las distorsiones de la ley de Coulomb son más drásticas, pues la distribución geométrica de los iones hace que pequeños campos eléctricos locales se acumulen hasta hacerse notables a nivel macroscópico. Este fenómeno se conoce como *polarización* eléctrica del medio. Comentaremos algo más sobre los efectos de la polarización, pero antes conviene introducir algunos conceptos adicionales.

En lugar de la fuerza que se ejercen dos partículas conviene hablar del *campo eléctrico* generado por una partícula: la *intensidad del campo eléctrico* en un punto del espacio es la fuerza que experimentaría una partícula con 1 culombio de carga (positiva) que estuviera situada en dicho punto. Si la carga Q que genera el campo está situada en la posición y, entonces el campo eléctrico en el punto x es

$$E(x) = \frac{1}{4\pi\epsilon}\frac{Q}{\|x-y\|^3}(x-y).$$

El vector de *inducción eléctrica* generada por una carga Q se define como

$$D(x) = \frac{1}{4\pi}\frac{Q}{\|x-y\|^3}(x-y).$$

De este modo, salvo por la constante ϵ_0, el vector D nos da el campo eléctrico que induce la carga Q sin tener en cuenta la acción del medio. Si el medio es homogéneo tenemos la relación $E = D/\epsilon$. En el seno de un elemento cristalino, donde la polarización se debe a una estructura geométrica simple, las propiedades del vector D siguen siendo válidas (pues hacen referencia a lo que ocurriría en el vacío) y el único cambio es que el campo eléctrico total es de la forma $E = D/\epsilon + P$, donde P es un vector de polarización que depende de la estructura del cristal. En medios más irregulares la relación entre la inducción D y el campo real E que aparece a causa de la influencia del medio puede ser mucho más compleja. En lo que sigue nos limitaremos a considerar el caso en que $E = D/\epsilon$.

Puesto que la fuerza total que actúa sobre una partícula es la suma vectorial de las fuerzas que actúan sobre la misma, el campo eléctrico generado por un número finito de partículas es la suma de los campos que genera cada una de ellas. Como en el caso de la gravitación, podemos generalizar la ley de Coulomb para determinar el campo eléctrico determinado por una distribución continua de cargas en el espacio. Llamemos ρ a la *densidad de carga,* es decir, la función

cuya integral en una región da la carga contenida en la misma. Entonces el campo eléctrico en un punto x viene dado por

$$E(x) = \frac{1}{4\pi\epsilon} \int_\Omega \frac{\rho(y)}{\|x-y\|^3}(x-y)\, dm(y).$$

Equivalentemente, la inducción eléctrica de la distribución de carga ρ es

$$D(x) = \frac{1}{4\pi} \int_\Omega \frac{\rho(y)}{\|x-y\|^3}(x-y)\, dm(y).$$

El teorema 10.19 garantiza que $E(x)$ y $D(x)$ están bien definidos. Más aún, el campo eléctrico es conservativo y deriva del potencial

$$V(x) = \frac{1}{4\pi\epsilon} \int_\Omega \frac{\rho(y)}{\|x-y\|}\, dm(y),$$

es decir, $E = -\nabla V$. La interpretación física de V es que el trabajo necesario para transportar una carga Q entre dos puntos es igual a Q por la diferencia de potencial entre los mismos. Si medimos este trabajo en Julios y la carga Q en culombios, la unidad de potencial —es decir, el Julio/culombio— recibe el nombre de *voltio*.

El potencial verifica la ecuación de Poisson: $\Delta V = -\rho/\epsilon$, o equivalentemente, $\operatorname{div} E = \rho/\epsilon$. En términos de la inducción eléctrica queda una relación más simple: $\operatorname{div} D = \rho$. El teorema de la divergencia nos da el teorema de Gauss:

> *El flujo de la inducción eléctrica D a través de una superficie cerrada es igual a la carga neta que ésta encierra.*

El hecho de que E y D sean campos conservativos nos da también la relación $\operatorname{rot} D = \operatorname{rot} E = 0$.

Ahora debemos resaltar una restricción importantísma, y es que todo lo anterior sólo es aplicable al caso en que la distribución de la carga eléctrica no varía con el tiempo.[2] Antes de entrar en el caso general debemos detenernos en el estudio de la magnetostática.

13.2 Magnetostática

La manifestación más conocida de la fuerza magnética la constituyen los imanes. Se trata de ciertos cuerpos que atraen al hierro y a otros metales. Hay imanes naturales como la magnetita, y también se puede convertir en un imán (durante un tiempo prolongado) a una pieza de hierro, sin más que frotarla con otro imán. Un imán, cualquiera que sea su forma, presenta siempre dos

[2]No es necesario tomar esto al pie de la letra. Por ejemplo, si una carga eléctrica pequeña se mueve en el seno de un campo eléctrico potente, su movimiento afecta tan poco al entorno que las leyes electrostáticas son aplicables.

polos, a los que se les llama Norte y Sur. El nombre no es aleatorio, sino que se debe a que la Tierra tiene propiedades magnéticas, de modo que una aguja imantada (suspendida de modo que pueda girar libremente) orienta sus polos en la dirección Norte-Sur. Los polos homólogos de dos imanes se repelen, los polos opuestos se atraen. En un principio se pensó en la existencia de "cargas magnéticas" similares a las eléctricas y que explicaran el magnetismo, pero existen fenómenos que no encajan en ese esquema. Por ejemplo, si se rompe una barra imantada, los dos trozos no resultan ser un polo norte y un polo sur aislados, sino que cada parte pasa a tener su propio polo norte y su propio polo sur.

La realidad es que las fuerzas magnéticas las provoca el movimiento de las cargas eléctricas, y a su vez afecta a las cargas eléctricas en movimiento. La atracción del hierro y los imanes se debe al movimiento microscópico de los electrones en los átomos. Cuantitativamente, la fuerza magnética puede describirse mediante un campo vectorial B, de modo que la fuerza magnética que actúa sobre una partícula con carga q que se mueve con velocidad v viene dada por la ley de Laplace:

$$F = qv \wedge B.$$

Así pues, la fuerza magnética es nula sobre partículas sin carga eléctrica o sobre cargas en reposo. Teniendo en cuenta además el campo eléctrico, la fuerza total que actúa sobre una partícula cargada es

$$F = q(E + v \wedge B).$$

Del mismo modo, los campos magnéticos los producen las cargas eléctricas en movimiento. Al igual que la ley de Coulomb que determina el campo eléctrico a partir de la densidad de carga sólo es válida cuando ésta es constante, el campo magnético está determinado por la llamada *ley de Biot y Savart* bajo el supuesto de que tanto la densidad de carga como la *densidad de corriente* sean constantes.

Para introducir el concepto de densidad de corriente necesitamos descomponer la densidad de carga como $\rho = \rho_+ - \rho_-$, donde ρ_+ es la densidad de carga positiva y ρ_- es la densidad de carga negativa. Similarmente, llamamos v_+ al campo de velocidades de las cargas positivas y v_- al campo de velocidades de las cargas negativas. Así, la densidad de corriente se define como

$$\vec{\imath} = \rho_+ v_+ - \rho_- v_-.$$

Según razonamos al hablar de fluidos, la interpretación de $\vec{\imath}$ es que su flujo a través de una superficie es la cantidad de carga eléctrica neta (positiva menos negativa) que la atraviesa por unidad de tiempo. En particular, $\vec{\imath}$ se mide en culombios por segundo y metro cuadrado. La unidad culombio/segundo recibe el nombre de *amperio,* luego la densidad de corriente $\vec{\imath}$ se mide en amperios por metro cuadrado. El flujo de $\vec{\imath}$ se mide en amperios.

La conservación de la carga eléctrica positiva y negativa (respectivamente) viene expresada por el equivalente de la ecuación de continuidad en la mecánica

de fluidos, que en nuestro caso es

$$\operatorname{div}\rho_+ v_+ + \frac{\partial \rho_+}{\partial t} = 0, \qquad \operatorname{div}\rho_- v_- + \frac{\partial \rho_-}{\partial t} = 0,$$

y restando ambas ecuaciones tenemos la ecuación que expresa la conservación de la carga eléctrica:

$$\operatorname{div}\vec{\imath} + \frac{\partial \rho}{\partial t} = 0.$$

La magnetostática estudia el campo magnético bajo la hipótesis de que ρ y $\vec{\imath}$ son constantes, en cuyo caso la ecuación anterior se reduce a

$$\operatorname{div}\vec{\imath} = 0.$$

En este contexto, la ley de Biot y Savart afirma que el campo magnético viene dado por

$$B(x) = \frac{\mu}{4\pi}\int_\Omega \frac{\vec{\imath}(y)\wedge(x-y)}{\|x-y\|^3}\,dm(y),$$

donde Ω es un abierto que contenga a todas las cargas y μ es una constante llamada *permeabilidad magnética.* La situación es la misma que en la electrostática: la constante μ depende del medio. La permeabilidad del vacío se representa por μ_0 y la ley de Biot y Savart vale también cuando las partículas están inmersas en un medio homogéneo sin más que cambiar μ_0 por otra constante $\mu = \mu_r\mu_0$, para una cierta *permeabilidad relativa* μ_r dependiente del medio. La permeabilidad en el vacío vale $\mu_0 = 4\pi\cdot 10^{-7}$ en las unidades usuales.[3]

Teniendo en cuenta que

$$\vec{\jmath}\,dm = \rho_+v_+\,dm - \rho_-v_-dm = v_+dq - v_-dq,$$

la ley de Biot y Savart afirma la contribución al campo magnético en un punto x de una partícula con carga dq que se mueve con velocidad v en el punto y es perpendicular a v y a $x-y$, directamente proporcional al seno del ángulo entre estos dos vectores y a su carga eléctrica, e inversamente proporcional al cuadrado de la distancia. No obstante esto no es cierto para una partícula puntual aislada, pues una partícula puntual en movimiento produce una densidad de carga variable, y nos salimos entonces del contexto de la magnetostática.

Conviene introducir la *inducción magnética*

$$H(x) = \frac{1}{4\pi}\int_\Omega \frac{\vec{\imath}(y)\wedge(x-y)}{\|x-y\|^3}\,dm(y).$$

De este modo, H determina el campo magnético que se generaría en ausencia del medio. En medios homogéneos tenemos la relación $B = \mu H$. Vamos a

[3]Es decir, tomando el Kilogramo, el metro, el segundo y el culombio como unidades de masa, longitud, tiempo y carga eléctrica. La arbitrariedad en la definición del culombio permite prefijar el valor de una de las constantes μ_0 o ϵ_0. La otra debe medirse. La definición actual de culombio hace que μ_0 sea exacta y ϵ_0 aproximada.

justificar que esta integral (o, equivalentemente, la que define a B) existe siempre, bajo el supuesto de que $\vec{\imath}$ es una función de clase C^2 y acotada en Ω. Unas simples comprobaciones nos dan que

$$\frac{\vec{\imath}(y) \wedge (x-y)}{\|x-y\|^3} = \nabla_x \frac{1}{\|x-y\|} \wedge \vec{\imath}(y) = \operatorname{rot}_x \frac{\vec{\imath}(y)}{\|x-y\|}.$$

Por consiguiente

$$H(x) = \frac{1}{4\pi} \operatorname{rot} \int_\Omega \frac{\vec{\imath}(y)}{\|x-y\|}\, dm(y).$$

Las componentes de la última integral son potenciales newtonianos, luego la integral existe y la que define a H también. Definimos el *potencial vectorial* del campo magnético como

$$A(x) = \frac{\mu}{4\pi} \int_\Omega \frac{\vec{\imath}(y)}{\|x-y\|}\, dm(y),$$

con lo que $B = \operatorname{rot} A$, lo que a su vez implica $\operatorname{div} B = 0$. Por otra parte, el teorema 10.19 nos da la relación $\Delta A = -\mu\vec{\imath}$.

Vamos a calcular el rotacional de H. Por la definición del laplaciano vectorial tenemos que

$$\mu \operatorname{rot} H = \operatorname{rot}\operatorname{rot} A = \nabla \operatorname{div} A - \Delta A = \nabla \operatorname{div} A + \mu\vec{\imath}.$$

Calculemos

$$\begin{aligned}
\operatorname{div} A &= \frac{\mu}{4\pi} \int_\Omega \operatorname{div}_x \frac{\vec{\imath}(y)}{\|x-y\|}\, dm(y) = \frac{\mu}{4\pi} \int_\Omega \vec{\imath}(y) \nabla_x \frac{1}{\|x-y\|}\, dm(y) \\
&= -\frac{\mu}{4\pi} \int_\Omega \vec{\imath}(y) \nabla_y \frac{1}{\|x-y\|}\, dm(y) \\
&= \frac{\mu}{4\pi} \int_\Omega \frac{\operatorname{div}_y \vec{\imath}(y)}{\|x-y\|}\, dm(y) - \frac{\mu}{4\pi} \int_\Omega \operatorname{div}_y \frac{\vec{\imath}(y)}{\|x-y\|}\, dm(y)
\end{aligned}$$

La primera integral es nula porque $\operatorname{div}_y \vec{\imath} = 0$ y, tomando Ω suficientemente grande como para que $\vec{\imath}$ sea nulo en $\partial\Omega$, el teorema de la divertencia implica que la segunda también lo es. Concluimos que $\operatorname{div} A = 0$ y, por consiguiente,

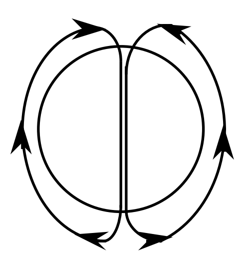
$$\operatorname{rot} H = \vec{\imath}.$$

En total tenemos las ecuaciones

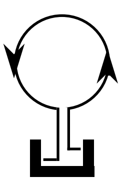
$$\begin{array}{ll}
\operatorname{div} D = \rho & \operatorname{rot} E = 0 \\
\operatorname{div} B = 0 & \operatorname{rot} H = \vec{\imath}
\end{array}$$

La última ecuación implica que el flujo de corriente a través de una superficie es igual a la circulación de H alrededor de su frontera.

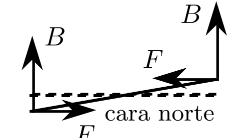
Terminamos la sección explicando brevemente la relación entre la teoría que acabamos de describir y los fenómenos magnéticos cotidianos (imanes, brújulas, etc.)



Ante todo, la Tierra se halla rodeada de un campo magnético cuyas causas no se comprenden completamente y dependen de la estructura de su núcleo. Las "líneas de fuerza" del campo magnético, es decir, sus curvas integrales, parten del polo sur, donde B es prácticamente vertical, se doblan rápidamente y ascienden hacia el polo norte, donde penetran de nuevo en la tierra (de hecho las líneas de fuerza son cerradas, de modo que descienden por el interior de la Tierra hasta volver al punto de partida). En un punto de la superficie no muy cercano a los polos el campo magnético terrestre puede considerarse constante y apunta en la dirección sur-norte.

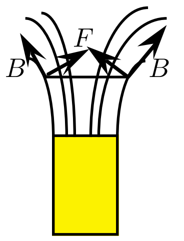


Consideremos un circuito eléctrico circular suspendido verticalmente de modo que pueda girar sobre un eje vertical. Llamemos cara Norte del circuito a la cara que muestra la figura, de modo que la corriente se ve circular en sentido antihorario. La cara opuesta será la cara Sur. Es fácil ver que el campo magnético terrestre hará girar el circuito hasta que su cara norte quede encarada hacia el polo norte terrestre. En tal caso la fuerza magnética en cada punto del circuito apunta hacia fuera del mismo, luego no mueve la espira.



Teóricamente también podría alcanzarse el equilibrio con la cara sur apuntando al norte, pero sería un equilibrio inestable, pues la más leve perturbación haría girar el circuito hasta que la cara norte apuntara al norte.

El mismo principio explica por qué se orientan las agujas imantadas. La imantación de la materia se debe a su estructura atómica. Un modelo simplista, pero que basta a nuestros propósitos, consiste en imaginar un imán como una porción de materia cuyos electrones giran todos en el mismo plano en el mismo sentido, de modo que su polo norte es la superficie mirándolo desde la cual el sentido de giro es antihorario. Entonces el argumento anterior se aplica a cada átomo del imán. Las fuerzas internas entre átomos contiguos se cancelan, pero no así las fuerzas en la superficie del imán, luego el resultado es que el campo magnético orienta al imán. Puede probarse que el campo magnético que genera un imán es similar al terrestre, pero las líneas de fuerza surgen del polo norte y vuelven a entrar en el imán por el polo sur (la Tierra es, pues, un imán con su polo norte en el polo sur geográfico y su polo sur en el polo norte geográfico).



Es claro que dos polos iguales de dos imanes se repelen y dos polos opuestos se atraen. En efecto, las líneas de fuerza que salen del polo norte de un imán se abren como las hojas de una palmera. Si identificamos cada átomo del otro polo norte con un circuito circular el resultado es una fuerza repulsiva. Si los polos son opuestos la fuerza resulta atractiva. La intensidad de la fuerza magnética depende del flujo de B a través de los polos del

imán. Argumentos similares explican que una brújula (horizontal) se desvía en presencia de una corriente eléctrica vertical. Es fácil ver que las líneas de fuerza del campo magnético de una corriente rectilínea son circunferencias alrededor del cable. La brújula se orienta según la dirección de B. Los imanes atraen al hierro y otros metales porque los átomos de hierro se orientan con facilidad según el campo magnético del imán, con lo que se convierten momentáneamente en imanes.

13.3 Las ecuaciones de Maxwell

En las secciones anteriores hemos descrito algunas de las leyes de la electricidad y el magnetismo, pero también hemos comentado que su validez está limitada a ciertas situaciones especiales. Fue J.C. Maxwell el primero que estableció (en 1863) las ecuaciones que caracterizan los campos eléctrico y magnético en condiciones generales, a partir de las cuales se deducen todas las leyes del electromagnetismo. Los hechos que hemos comentado en la sección anterior muestran que los campos eléctrico y magnético no son entidades independientes, sino que están muy relacionados, por lo que a menudo se habla del "campo electromagnético". El punto más delicado es establecer con exactitud la dependencia entre ambos. Así, la ecuación magnetostática $\operatorname{rot} H = \vec{\imath}$ se generaliza a la ecuación

$$\operatorname{rot} H = \vec{\imath} + \frac{\partial D}{\partial t},$$

que muestra cómo la variación del campo eléctrico produce un campo magnético. Del mismo modo, sucede que la ecuación electrostática $\operatorname{rot} E = 0$ es incompleta, y en el caso general debe modificarse para reflejar que las variaciones de campos magnéticos también pueden producir campos eléctricos. Este fenómeno fue descubierto por Faraday investigando con bobinas, es decir, circuitos en forma espiral. En la sección anterior hemos visto que un campo magnético puede hacer girar un circuito en forma de espira. Recíprocamente, Faraday descubrió que si hacemos girar un circuito en forma de espira en el seno de un campo magnético, éste inducirá una corriente en el circuito. Debidamente generalizada, la ley de Faraday puede enunciarse como sigue:

> *Si C es un circuito que rodea una superficie S, entonces la circulación de E a lo largo de C coincide con el opuesto de la variación del flujo de B a través de S.*

Simbólicamente:

$$\int_C E\, d\vec{r} = -\frac{\partial}{\partial t} \int_S d\Phi(B).$$

El signo expresa una simple cuestión de orientación: fijado el sentido en que el flujo se considera positivo, un aumento del mismo induce sobre C una corriente en sentido horario. Esta ley explica el funcionamiento de los alternadores: cuando un circuito en forma de espira es hecho girar en un campo magnético

constante, el flujo que lo atraviesa pasa de un valor máximo F hasta el valor $-F$ de forma periódica, lo que provoca una corriente alterna en el circuito.

Por el teorema de Stokes tenemos

$$\int_S d\Phi(\operatorname{rot} E) = \int_S d\Phi\left(-\frac{\partial B}{\partial t}\right),$$

y como la igualdad se ha de dar para toda superficie S es claro que ha de ser

$$\operatorname{rot} E = -\frac{\partial B}{\partial t}.$$

Con esto tenemos la última de las *ecuaciones de Maxwell.* En total son las siguientes:

$$\begin{aligned} \operatorname{div} D &= \rho \qquad & \operatorname{div} B &= 0 \\ \operatorname{rot} E &= -\frac{\partial B}{\partial t} \qquad & \operatorname{rot} H &= \vec{\imath} + \frac{\partial D}{\partial t}. \end{aligned}$$

La elección de D o E y de B o H no se debe a razones estéticas (para eliminar constantes), sino que de este modo las ecuaciones valen para medios arbitrarios. En medios homogéneos debemos completarlas con las relaciones

$$D = \epsilon E, \quad B = \mu H.$$

Ninguno de los argumentos de las secciones anteriores puede considerarse una demostración de las ecuaciones de Maxwell, pues todos ellos partían de hipótesis restrictivas sobre invarianza en el tiempo. La teoría del electromagnetismo de Maxwell postula la validez general de estas cuatro ecuaciones y deduce de ellas todas las propiedades del electromagnetismo, incluidos los resultados de las secciones anteriores. Por ejemplo, tomando la divergencia en la cuarta ecuación queda

$$0 = \operatorname{div}\operatorname{rot} H = \operatorname{div} \vec{\imath} + \frac{\partial}{\partial t}\operatorname{div} D,$$

y usando la primera ecuación llegamos a

$$\operatorname{div} \vec{\imath} = -\frac{\partial \rho}{\partial t},$$

es decir, tenemos la ecuación de continuidad, que expresa la conservación de la carga eléctrica. Por otra parte, el teorema de Gauss se sigue inmediatamente de la primera ecuación mediante el teorema de la divergencia.

Los potenciales De $\operatorname{div} B = 0$ se sigue la existencia del potencial vectorial magnético A, de modo que $B = \operatorname{rot} A$. La tercera ecuación de Maxwell muestra que en general el campo eléctrico no es conservativo, pero sustituyendo el potencial magnético queda

$$\operatorname{rot}\left(E + \frac{\partial A}{\partial t}\right) = 0,$$

con lo que existe un potencial escalar V tal que

$$E = -\nabla V - \frac{\partial A}{\partial t}. \tag{13.1}$$

En el caso en que A no dependa de t tenemos el potencial electrostático que ya conocemos. Las ecuaciones de Maxwell permiten expresar los potenciales V y A en términos de la distribución de carga ρ y la densidad de corriente $\vec{\imath}$. El primer paso es encontrar ecuaciones diferenciales que contengan una sola de las incógnitas V y A. Para ello sustituimos la ecuación anterior en la primera ecuación de Maxwell:

$$-\Delta V - \frac{\partial}{\partial t}\operatorname{div} A = \frac{\rho}{\epsilon}. \tag{13.2}$$

Por otro lado, sustituyendo (13.1) y $B = \operatorname{rot} A$ en la cuarta ecuación de Maxwell resulta

$$\operatorname{rot}\operatorname{rot} A = \mu\vec{\imath} + \mu\epsilon\left(-\nabla\frac{\partial V}{\partial t} - \frac{\partial^2 A}{\partial t^2}\right).$$

Ahora usamos la definición de laplaciano vectorial:

$$\nabla\operatorname{div} A - \Delta A = \mu\vec{\imath} - \mu\epsilon\nabla\frac{\partial V}{\partial t} - \mu\epsilon\frac{\partial^2 A}{\partial t^2} \tag{13.3}$$

Las ecuaciones (13.2) y (13.3) se pueden simplificar notablemente si tenemos en cuenta que A está determinado únicamente por su rotacional, y el teorema 11.21 nos permite elegir su divergencia. Una buena elección es

$$\operatorname{div} A = -\mu\epsilon\frac{\partial V}{\partial t}. \tag{13.4}$$

Esta ecuación se conoce como *condición de Lorentz* y tiene una interpretación en la teoría de la relatividad. Sustituyéndola en las ecuaciones que hemos obtenido resultan dos fórmulas simétricas:

$$\Delta V - \mu\epsilon\frac{\partial^2 V}{\partial t^2} = -\frac{\rho}{\epsilon} \tag{13.5}$$

$$\Delta A - \mu\epsilon\frac{\partial^2 A}{\partial t^2} = -\mu\vec{\imath} \tag{13.6}$$

Estas ecuaciones determinan los potenciales V y A a partir de ρ e $\vec{\imath}$, y a su vez los potenciales determinan los campos E y B a través de las ecuaciones

$$B = \operatorname{rot} A, \quad E = -\nabla V - \frac{\partial A}{\partial t}.$$

Los campos E y B satisfacen ecuaciones similares a (13.5) y (13.6). Para obtener la de E tomamos gradientes en la primera ecuación de Maxwell y aplicamos la definición del laplaciano vectorial:

$$\Delta E + \operatorname{rot}\operatorname{rot} E = \frac{1}{\epsilon}\nabla\rho.$$

Aplicamos la tercera ecuación y después la cuarta:

$$\Delta E - \frac{\partial}{\partial t}\operatorname{rot} B = \frac{1}{\epsilon}\nabla\rho,$$

$$\Delta E - \mu\left(\frac{\partial\vec{\imath}}{\partial t} + \frac{\partial^2 D}{\partial t^2}\right) = \frac{1}{\epsilon}\nabla\rho.$$

De aquí llegamos a

$$\Delta E - \mu\epsilon\frac{\partial^2 E}{\partial t^2} = \frac{1}{\epsilon}\nabla\rho + \mu\frac{\partial\vec{\imath}}{\partial t}. \tag{13.7}$$

Un razonamiento similar nos da

$$\Delta H - \mu\epsilon\frac{\partial^2 H}{\partial t^2} = -\operatorname{rot}\vec{\imath}. \tag{13.8}$$

Las ecuaciones en derivadas parciales (13.5), (13.6), (13.7) y (13.8) se diferencian únicamente en el término independiente (el que no contiene a la incógnita V, A, E o H). Ello hace que su resolución pueda ser estudiada en general, cosa que hacemos en la sección siguiente. Notemos, no obstante, que si las magnitudes son invariantes con el tiempo todas ellas son ecuaciones de Poisson, que ya sabemos resolver, y nos llevan a los resultados de las secciones anteriores.

Energía electromagnética Apliquemos la relación (10.11) a los campos E y H, es decir:

$$\operatorname{div}(E\wedge H) = (\operatorname{rot} E)H - E(\operatorname{rot} H).$$

Si lo aplicamos a los campos E y H, las ecuaciones de Maxwell implican

$$\operatorname{div}(E\wedge H) = -H\,\frac{\partial B}{\partial t} - E\,\frac{\partial D}{\partial t} - E\vec{\imath} = -\frac{\partial}{\partial t}\frac{\mu H^2}{2} - \frac{\partial}{\partial t}\frac{\epsilon E^2}{2} - E\vec{\imath}.$$

Fijado un volumen Ω, podemos aplicar el teorema de la divergencia:

$$\frac{\partial}{\partial t}\int_\Omega\left(\frac{\mu H^2}{2} + \frac{\epsilon E^2}{2}\right)dm + \int_\Omega E\vec{\imath}\,dm = -\int_{\partial\Omega} d\Phi(E\wedge H) \tag{13.9}$$

Observemos que si sobre un objeto de masa m que se mueve con velocidad v actúa una fuerza F, su energía cinética es $(1/2)mv^2$, luego la variación de esta energía es $mva = vF$. En nuestro caso, si llamamos v a la velocidad de las cargas en cada punto (y recordando que dm es el elemento de volumen), se cumple $E\vec{\imath}\,dm = E\rho v\,dm = Ev\,dq = (E + v\wedge B)v\,dq = vdF$, luego el segundo término del primer miembro de la igualdad anterior es el aumento de energía cinética del fluido eléctrico producido por el campo electromagnético.

Esto nos lleva a definir la *energía potencial electromagnética* acumulada en un volumen Ω como

$$\mathcal{E} = \int_\Omega\left(\frac{\mu H^2}{2} + \frac{\epsilon E^2}{2}\right)dm$$

De este modo, la relación anterior es una ley de conservación: si tomamos Ω suficientemente grande como para que los campos sean nulos en su frontera, tenemos que la energía total (cinética más potencial) permanece constante. Más en general, el vector $P = E \wedge H$ recibe el nombre de *vector de Poynting,* y su flujo a través de una superficie nos da la energía electromagnética que sale de ella por unidad de tiempo. La variación de la energía (o el trabajo realizado por unidad de tiempo) se mide en *vatios.* Un vatio es simplemente un Julio por segundo. Las unidades del vector de Poynting son, pues, vatios por metro cuadrado.

La ley de Ohm Una de las aplicaciones más importantes del electromagnetismo lo constituyen los circuitos eléctricos, formados por una red de cables conductores por los que se hace circular una corriente eléctrica. Precisemos estos conceptos. Un conductor es una sustancia cuya estructura atómica permite que sus electrones pasen con facilidad de un átomo a otro. Los mejores conductores son los metales. Los electrones exteriores de los átomos metálicos no permanecen vinculados a ningún átomo en particular, sino que forman una nube compartida por todos los átomos. Por ello un trozo de metal puede considerarse como una molécula gigante. Cuando aplicamos un campo eléctrico constante a un electrón en el vacío éste emprende un movimiento con aceleración constante. No les sucede lo mismo a los electrones libres de un conductor, pues en presencia del campo van chocando de átomo en átomo, siendo absorbidos y reemitidos, lo que les provoca un retardo considerable y su movimiento a nivel macroscópico puede considerarse uniforme (sin aceleración). Por otro lado, los choques hacen vibrar los átomos, lo que macroscópicamente se traduce en un calentamiento del conductor. Existe una fórmula sencilla avalada por la experiencia que permite tratar matemáticamente estos fenómenos y que se conoce como *ley de Ohm.* En su forma general puede enunciarse así: si aplicamos un campo eléctrico E a un conductor, la densidad de corriente en cada punto vendrá dada por $\vec{\imath} = \sigma E$, donde σ es una constante que depende del conductor y recibe el nombre de *conductividad.* La conductividad permite hablar de buenos y malos conductores. Un conductor es mejor cuanto mayor es su conductividad, pues significa que un mismo campo aplicado produce una corriente más intensa.

Observar que el vector E, y por lo tanto $\vec{\imath}$ apunta hacia donde se movería una carga positiva situada en la posición considerada pero, como las cargas que se mueven son negativas, en realidad $\vec{\imath}$ tiene el sentido contrario al flujo de electrones. Desde un punto de vista matemático, una corriente de cargas negativas es equivalente a una hipotética corriente de cargas positivas que circula en sentido contrario, y en muchas ocasiones es más claro pensarlo así.

Podemos distinguir entre *corrientes de conducción,* en las que los electrones que se desplazan son sustituidos por otros, de modo que la carga neta es siempre nula, y las *corrientes de convección,* en las que se produce una acumulación de electrones en unos puntos y un defecto en otros. Cuando las corrientes son de conducción podemos suponer que $\rho = 0$, reflejando el hecho de que cualquier integral de ρ en un volumen macroscópico será nula. La existencia de cargas en el medio ya está implícita en las constantes ϵ y μ específicas del conductor.

Veamos algunas consecuencias de la ley de Ohm aplicada a un circuito eléctrico. Podemos representar un tramo de cable eléctrico como una curva diferenciable $\gamma(s)$, aunque en realidad el cable será un tubo, digamos cilíndrico, que rodea a γ. Sea k su sección (que podemos considerar constante o función del parámetro s). El conductor está rodeado de un medio aislante, de modo que sólo la parte de E en la dirección T tangente a γ produce corriente. Así pues, la ley de Ohm debe ser adaptada a $\vec{\imath} = (ET)T$. Con esto suponemos implícitamente que la densidad de corriente $\vec{\imath}$ es la misma en cada sección transversal del conductor. Definimos la *intensidad de corriente* como $I = kT\vec{\imath}$, es decir, I es el flujo de $\vec{\imath}$ a través de la sección de cable en un punto dado o, lo que es lo mismo, la cantidad de carga (positiva) que la atraviesa por unidad de tiempo.

Definimos la *caída de tensión* entre los extremos del cable como la circulación de E a través del mismo, o sea,

$$dV = ET\,ds = \frac{\vec{\imath}\,T}{\sigma}\,ds = \frac{I}{k\sigma}\,ds.$$

Así pues, V es el trabajo que realiza el campo eléctrico para transportar un culombio (positivo) de un extremo a otro del cable. Si recorremos el cable en el sentido de la corriente positiva (es decir, en sentido opuesto a la corriente real) V ha de ser positivo. En ausencia de campos magnéticos variables el campo eléctrico es conservativo y V es la pérdida de energía potencial que sufre la unidad de carga al recorrer el cable. En cualquier caso V debería coincidir con la energía cinética que el campo comunica a la carga, pero según hemos comentado ésta no gana ninguna energía cinética, pues su velocidad media es constante. En realidad V es la cantidad de calor que se genera por cada carga positiva que se transporta a lo largo del cable. Notemos que V se mide en voltios.

Definimos la *resistencia* del cable como la integral de $dR = ds/k\sigma$, que es una constante asociada al conductor. Si éste es homogéneo es simplemente $R = L/k\sigma$, donde L es la longitud del cable, k su sección y σ su conductividad. La unidad de resistencia es el ohmio. Notemos que la caída de tensión es $dV = IdR$. Si la intensidad es constante queda simplemente $V = IR$, que es la forma habitual de la ley de Ohm.

Vamos a calcular el calor P que genera la corriente eléctrica por unidad de tiempo. Consideremos un elemento de cable de longitud ds. Sea dt el tiempo que la corriente tarda en recorrerlo. Durante este tiempo, la carga que sale del elemento de cable es exactamente la carga que contiene, luego ésta es $I\,dt$. Dicha carga recorre una distancia ds en el tiempo dt, lo cual le supone al campo un trabajo $I\,dV dt$. Así pues, $dP = I\,dV = I^2\,dR$. Si la intensidad es constante queda $P = I^2R = V^2/R$, que es la *ley de Joule.*

Observemos que $dP = I\,dV = k(T\vec{\imath})(ET)\,ds = E\vec{\imath}\,dm$, con lo que el segundo término del primer miembro de la ecuación (13.9) es ahora el calor desprendido por el cable por unidad de tiempo.

13.4 La ecuación de ondas

En esta sección abordamos el problema matemático de resolver las ecuaciones en derivadas parciales que determinan los potenciales y los campos electromagnéticos. Se define el operador *d'alembertiano* (tridimensional) de constante $v > 0$ como el que a cada función u en las variables x, y, z, t de clase C^2 le hace corresponder la función

$$\Box_v u = \frac{\partial^2 u}{\partial t^2} - v^2 \Delta u = \frac{\partial^2 u}{\partial t^2} - v^2 \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 u}{\partial z^2} \right)$$

Si u es una función vectorial (con valores en $\mathbb{R}^3$) su d'alembertiano se define análogamente usando el laplaciano vectorial en lugar del escalar. Es claro entonces que $\Box_v u = (\Box_v u_1, \Box_v u_2, \Box_v u_3)$.

Similarmente se definen el d'alembertiano bidimensional (para funciones de x, y, t) y el unidimensional (para funciones de x, t). En todas las aplicaciones, las variables x, y, z representan una posición en el espacio y t representa al tiempo. La función $u(x, y, z, t)$ representa la *intensidad* de una magnitud física que varía con el tiempo de forma distinta en cada punto.

Cualquier ecuación en derivadas parciales de la forma $\Box_v u = w$ se llama *ecuación de D'Alembert* o *ecuación de ondas,* debido a que sus soluciones representan procesos ondulatorios tales como la vibración de una cuerda o una membrana, la transmisión del sonido en un fluido o —lo que nos interesará especialmente en este capítulo— la luz.

Si llamamos $c = 1/\sqrt{\epsilon\mu}$ las ecuaciones (13.5), (13.6), (13.7) y (13.8) equivalen a las siguientes ecuaciones de ondas:

$$\Box_c V = \frac{c^2 \rho}{\epsilon} \tag{13.10}$$

$$\Box_c A = \mu c^2 \vec{\imath} \tag{13.11}$$

$$\Box_c E = -\frac{c^2}{\epsilon} \nabla \rho - c^2 \mu \frac{\partial \vec{\imath}}{\partial t} \tag{13.12}$$

$$\Box_c H = c^2 \operatorname{rot} \vec{\imath}. \tag{13.13}$$

La constante c depende del medio. Una simple comprobación muestra que las unidades de ϵ son C^2/Nm^2 y las de μ son Ns^2/C^2, con lo que las unidades de c son metros por segundo, es decir, corresponde a una velocidad. En el vacío vale

$$c \approx \sqrt{\frac{36\pi \cdot 10^9}{4\pi \cdot 10^{-7}}} = 3 \cdot 10^8 \text{ m/s} = 300.000 \text{ Km/s}.$$

Ésta es aproximadamente la velocidad de la luz. Maxwell fue el primero en explicar la naturaleza de la luz en términos de la teoría electromagnética. Antes de entrar en ello debemos investigar las soluciones de la ecuación de ondas, que es lo que nos ocupa en esta sección.

Ecuación homogénea unidimensional Nos ocupamos en primer lugar de la ecuación

$$\frac{\partial^2 u}{\partial t^2} - v^2 \frac{\partial^2 u}{\partial x^2} = 0,$$

es decir, de la ecuación de ondas unidimensional y homogénea. Para resolverla hacemos el cambio $\xi = x + vt$, $\eta = x - vt$. Según el convenio habitual, llamamos $u(\xi, \eta)$ a la composición de $u(x,t)$ con el cambio inverso. Aplicando la regla de la cadena tenemos

$$\begin{aligned}
\frac{\partial u}{\partial t} &= v\frac{\partial u}{\partial \xi} - v\frac{\partial u}{\partial \eta} \\
\frac{\partial u}{\partial x} &= \frac{\partial u}{\partial \xi} + \frac{\partial u}{\partial \eta} \\
\frac{\partial^2 u}{\partial t^2} &= v^2\frac{\partial^2 u}{\partial \xi^2} - 2v^2\frac{\partial^2 u}{\partial \xi \partial \eta} + v^2\frac{\partial^2 u}{\partial \eta^2} \\
\frac{\partial^2 u}{\partial x^2} &= \frac{\partial^2 u}{\partial \xi^2} + 2\frac{\partial^2 u}{\partial \xi \partial \eta} + \frac{\partial^2 u}{\partial \eta^2}
\end{aligned}$$

Al sustituir estas derivadas en la ecuación de ondas ésta se reduce a

$$\frac{\partial^2 u}{\partial \xi \partial \eta} = 0. \tag{13.14}$$

Esto significa que la derivada de u respecto de η ha de ser una función $f(\xi)$ que no dependa de η. Por consiguiente

$$u = \int_{\eta_0}^{\eta} f(s)\, ds + f_1(\xi) = f_1(\xi) + f_2(\eta).$$

Recíprocamente, cualquier función de la forma $u(\xi, \eta) = f_1(\xi) + f_2(\eta)$, para ciertas funciones f_1 y f_2 de clase C^2, es solución de (13.14). Deshaciendo el cambio de variable, concluimos que las soluciones de la ecuación de ondas son todas las funciones de la forma

$$u(x, y) = f_1(x + vt) + f_2(x - vt). \tag{13.15}$$

Ahora veamos que la ecuación tiene solución única si especificamos arbitrariamente los valores de u y de su derivada respecto a t en un instante dado, es decir, vamos a probar que el problema siguiente tiene solución única $u(x,t)$:

$$\left.\begin{aligned}
\frac{\partial^2 u}{\partial t^2} &= v^2\ \frac{\partial^2 u}{\partial x^2} \\
u(x, 0) &= \phi(x) \\
\frac{\partial u}{\partial t}(x, 0) &= \psi(x)
\end{aligned}\right\}$$

donde ϕ es una función de clase C^2 y ψ de clase C^1.

Para que una solución u dada por (13.15) satisfaga las condiciones iniciales ha de cumplir

$$\begin{aligned} f_1(x) + f_2(x) &= \phi(x) \\ v f_1'(x) - v f_2'(x) &= \psi(x) \end{aligned}$$

De la segunda ecuación deducimos

$$f_1(x) - f_2(x) = \frac{1}{v}\int_{x_0}^{x} \psi(s)\, ds + C,$$

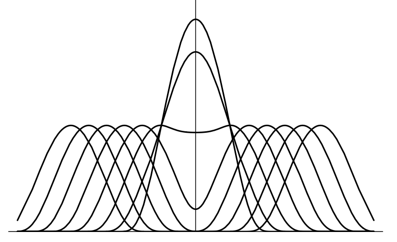
y ahora resolviendo el sistema lineal de incógnitas $f_1(x)$ y $f_2(x)$ obtenemos

$$\begin{aligned} f_1(x) &= \frac{1}{2}\phi(x) + \frac{1}{2v}\int_{x_0}^{x} \psi(s)\, ds + \frac{C}{2} \\ f_2(x) &= \frac{1}{2}\phi(x) - \frac{1}{2v}\int_{x_0}^{x} \psi(s)\, ds - \frac{C}{2} \end{aligned}$$

Entonces (13.15) se convierte en

$$u(x,t) = \frac{\phi(x+vt) + \phi(x-vt)}{2} + \frac{1}{2v}\int_{x-vt}^{x+vt} \psi(s)\, ds. \qquad (13.16)$$

Esta fórmula indica que el valor $u(x,t)$ es el promedio de los valores que tomaba u para $t = 0$ en los puntos $x \pm vt$ más el promedio de la velocidad inicial en el intervalo $[x - vt, x + vt]$ multiplicado por t. Recíprocamente, la situación inicial en un punto x_0 afecta en cada instante t a los puntos que distan de x_0 un máximo de vt. Así pues, la constante v es la velocidad con que se expande el radio de influencia del estado inicial de cada punto sobre el estado de u.



La figura muestra los valores de u para tiempos distintos a partir del estado inicial determinado por una función ϕ en forma de cresta (la gráfica más alta) y velocidades nulas ($\psi = 0$). Vemos cómo la cresta se va achatando por el centro hasta dividirse en dos crestas menores, una que se mueve hacia la izquierda y otra hacia la derecha, ambas con velocidad v. Desde el punto de vista de una posición fija alejada de la perturbación inicial, la función u comienza siendo nula, en un momento dado (tras el tiempo necesario para que el frente de onda llegue al punto) u comienza a crecer hasta llegar a un valor máximo y luego vuelve a decrecer hasta hacerse nula de nuevo. De existir una velocidad inicial ψ, digamos de soporte compacto y con integral no nula, su efecto sobre los puntos alejados es permanente, en el sentido de que para tiempos t suficientemente grandes u terminará tomando el valor de dicha integral.

Ecuación homogénea tridimensional A continuación nos ocupamos del problema

$$\left.\begin{array}{rcl} \dfrac{\partial^2 u}{\partial t^2} = v^2 \Delta u & & \\ u(x,0) & = & \phi(x) \\ \dfrac{\partial u}{\partial t}(x,0) & = & \psi(x) \end{array}\right\}$$

donde ahora $x \in \mathbb{R}^3$. Es natural conjeturar que una solución de esta ecuación vendrá dada por una fórmula que generalice de algún modo a (13.16). El primer término de (13.16) es la media de los valores iniciales de u en los puntos $x \pm vt$, por lo que es razonable conjeturar que en el caso tridimensional aparecerá la media del estado inicial de u sobre la esfera de centro x y radio vt. Por ello conviene introducir la *media esférica* de una función $u : \mathbb{R}^3 \longrightarrow \mathbb{R}$ como la función

$$M(u)(x,r) = \frac{1}{4\pi r^2} \int_{\|y-x\|=r} u(y)\, d\sigma(y) = \frac{1}{4\pi} \int_{\|\xi\|=1} u(x + r\xi)\, d\sigma(\xi),$$

donde σ representa la medida de Lebesgue en la esfera de radio r en la primera integral y en la esfera de radio 1 en la segunda.

Para justificar el cambio de variable consideramos la aplicación $f(\xi) = x+r\xi$ y observamos que el plano tangente a la esfera de centro x y radio r en un punto $x+r\xi$ coincide con el plano tangente a la esfera de centro 0 y radio 1 en el punto ξ, y para todo par de vectores u, v en dicho plano.

$$f^\sharp(d\sigma_r)(\xi)(u,v) = d\sigma_r(x + r\xi)(ru, rv) = r^2 d\sigma_r(x + r\xi)(u,v) = r^2 d\sigma_1(\xi)(u,v),$$

pues $d\sigma_r(x + r\xi)(u,v)$ y $d\sigma_1(\xi)(u,v)$ son ambos el área del paralelogramo de lados u y v. Por consiguiente $f^\sharp(d\sigma_r) = r^2 d\sigma_1$ y basta aplicar el teorema 9.22.

Observemos que la segunda integral en la definición de $M(u)$ está definida para todo valor de r, no necesariamente positivo. De hecho es fácil ver que $M(u)(x,r) = M(u)(x,-r)$. Además $M(u)(x,0) = f(x)$.

La primera integral conecta directamente con el problema que estamos estudiando, pero la segunda es más fácil de manejar. Por ejemplo, nos permite calcular la derivada

$$\frac{\partial}{\partial r} M(u)(x,r) = \frac{1}{4\pi} \int_{\|\xi\|=1} \nabla u(x + r\xi)\xi\, d\sigma(\xi).$$

Notemos que ξ coincide con el vector normal unitario a la esfera en el punto ξ, luego la última integral es el flujo del campo $\nabla u(x + r\xi)$. Podemos aplicar el teorema de la divergencia y resulta:

$$\begin{aligned} \frac{\partial}{\partial r} M(u)(x,r) &= \frac{r}{4\pi} \int_{\|\xi\|<1} \Delta u(x + r\xi)\, dm(\xi) \qquad (13.17) \\ &= \frac{1}{4\pi r^2} \int_{\|y-x\|<r} \Delta u(y)\, dm(y). \end{aligned}$$

Supongamos ahora que $u(x,t)$ es una solución de la ecuación de ondas. Entonces las medias $M(u)(x,r,t)$ tienen a t como parámetro. Sustituyendo Δu por el valor que da la ecuación de ondas podemos continuar el cálculo anterior (para $r > 0$):

$$\frac{\partial}{\partial r} M(u)(x,r,t) = \frac{1}{4\pi r^2} \int_{\|y-x\|<r} \frac{1}{v^2} \frac{\partial^2 u}{\partial t^2}(y,t)\, dm(y)$$

$$= \frac{1}{4\pi r^2 v^2} \int_0^r \frac{\partial^2}{\partial t^2} \int_{\|y-x\|=\rho} u(y,t)\, d\sigma(y) d\rho = \frac{1}{r^2 v^2} \int_0^r \rho^2 \frac{\partial^2 M(u)}{\partial t^2}(x,\rho,t)\, d\rho.$$

Ahora multiplicamos por r^2 y derivamos respecto a r:

$$\frac{\partial}{\partial r}\left(r^2 \frac{\partial M(u)}{\partial r} \right) = \frac{r^2}{v^2} \frac{\partial^2 M(u)}{\partial t^2}.$$

Claramente entonces

$$\frac{1}{r} \frac{\partial^2 r M(u)}{\partial r^2} = \frac{1}{r^2} \frac{\partial}{\partial r}\left(r^2 \frac{\partial M(u)}{\partial r} \right) = \frac{1}{v^2} \frac{\partial^2 M(u)}{\partial t^2}.$$

Si llamamos $\overline{M(u)} = rM(u)$ tenemos

$$\frac{\partial^2 \overline{M(u)}}{\partial r^2} = \frac{1}{v^2} \frac{\partial^2 \overline{M(u)}}{\partial t^2},$$

es decir, la función $\overline{M(u)}(x,r,t)$, para un x fijo, cumple la ecuación de ondas unidimensional. Vamos a considerar a t como variable espacial (la que en el apartado anterior era x) y a r como variable temporal (la que en el apartado anterior era t). Entonces la constante de la ecuación de ondas es $1/v$ y según el apartado anterior sabemos que

$$\overline{M(u)}(x,r,t) = f_1(x, t + r/v) + f_2(x, t - r/v),$$

pero tenemos $\overline{M(u)}(x,0,t) = 0$, luego $f_1(x,t) + f_2(x,t) = 0$. Así pues, la solución depende de una única función $f(x,t) = f_1(x,t) = -f_2(x,t)$ y se cumple

$$\overline{M(u)}(x,r,t) = f(x, t + r/v) - f(x, t - r/v),$$

luego

$$M(u)(x,r,t) = \frac{1}{r}\big(f(t + r/v) - f(t + r/v) \big).$$

Ahora trataremos de recuperar $u(x,t)$ a partir de las medias $M(u)(x,r,t)$. Para ello notamos que $u(x,t) = M(u)(x,0,t) = (2/v) f'(x,t)$. Por otra parte

$$\frac{\partial r M(u)}{\partial r} + \frac{1}{v} \frac{\partial r M(u)}{\partial t} = \frac{2}{v} f'(x, t + r/v),$$

con lo que

$$u(x,t) = \frac{2}{v} f'(x,t) = \frac{\partial r M(u)}{\partial r}(x, vt, 0) + \frac{1}{v} \frac{\partial r M(u)}{\partial t}(x, vt, 0).$$

Para derivar respecto de r podemos hacer primero $t = 0$, y así

$$\frac{\partial rM(u)}{\partial r}(x, vt, 0) = \frac{\partial}{\partial r}\left(\frac{1}{4\pi r}\int_{\|y-x\|=r} u(y, 0)\, d\sigma(y)\right)\bigg|_{(x,vt)}$$

$$= \frac{\partial rM(\phi)}{\partial r}(x, vt) = \frac{\partial}{\partial t}\big(tM(\phi)(x, vt)\big).$$

Por otra parte,

$$\frac{\partial rM(u)}{\partial t}(x, vt, 0) = \left(\frac{1}{4\pi r}\int_{\|y-x\|=r} \frac{\partial u}{\partial t}(y, t)\, d\sigma(y)\right)\bigg|_{(x,vt,0)}$$

$$= \left(\frac{1}{4\pi r}\int_{\|y-x\|=r} \psi(y)\, d\sigma(y)\right)\bigg|_{(x,vt)} = \big(rM(\psi)\big)(x, vt) = vtM(\psi)(x, vt).$$

En total

$$u(x, t) = \frac{\partial}{\partial t}\big(tM(\phi)(x, vt)\big) + tM(\psi)(x, vt). \tag{13.18}$$

Explícitamente:

$$u(x, t) = \frac{\partial}{\partial t}\left(\frac{1}{4\pi v^2 t}\int_{\|y-x\|=vt} \phi(y)\, d\sigma(y)\right) + \frac{1}{4\pi v^2 t}\int_{\|y-x\|=vt} \psi(y)\, d\sigma(y).$$

Esto justifica la unicidad de la solución. Para probar la existencia hemos de ver que, para cualquier par de funciones ϕ de clase C^3 y ψ de clase C^2, la función dada por (13.18) satisface la ecuación de ondas con las condiciones iniciales ϕ y ψ. Comencemos por éstas:

$$\begin{aligned}
u(x, 0) &= M(\phi)(x, 0) + \left(t\frac{\partial}{\partial t}M(\phi)(x, vt)\right)\bigg|_{t=0} + 0 = \phi(x).\\
\frac{\partial u}{\partial t}(x, 0) &= \left(2\frac{\partial}{\partial t}M(\phi)(x, vt) + t\frac{\partial^2}{\partial t^2}M(\phi)(x, vt)\right)\bigg|_{(x,0)}\\
&+ \left(M(\psi)(x, vt) + t\frac{\partial}{\partial t}M(\psi)(x, vt)\right)\bigg|_{(x,0)} = \psi(x).
\end{aligned}$$

Hemos usado (13.17) para probar que

$$\frac{\partial}{\partial t}M(\phi)(x, vt)\bigg|_{(x,0)} = v\frac{\partial M(\phi)}{\partial r}(x, 0) = 0.$$

La suma de dos soluciones de la ecuación de ondas es también una solución, luego basta probar que los dos términos de (13.18) satisfacen la ecuación de ondas. Más aún, la derivada respecto de t de una solución es también solución,

luego basta ver que $tM(\psi)(x,vt)$ satisface la ecuación de ondas. Según (13.17) tenemos

$$\frac{\partial M(\psi)(x,vt)}{\partial t} = v\frac{\partial M(\psi)}{\partial r}(x,vt) = \frac{1}{4\pi vt^2}\int_{\|y-x\|<vt} \Delta\psi(y)\,dm(y)$$

$$= \frac{1}{4\pi vt^2}\Delta\int_0^{vt}\int_{\|y-x\|=\tau}\psi(y)\,d\sigma(y)d\tau$$

$$= \frac{1}{4\pi t^2}\Delta\int_0^{t}\int_{\|y-x\|=v\tau}\psi(y)\,d\sigma(y)d\tau = \frac{v^2}{t^2}\Delta\int_0^t \tau^2 M(\psi)(x,v\tau)\,d\tau.$$

Multiplicamos por t^2 y derivamos respecto de t:

$$\frac{\partial}{\partial t}\left(t^2\,\frac{\partial M(\psi)(x,vt)}{\partial t}\right) = v^2\Delta\big(t^2 M(\psi)(x,vt)\big)$$

Entonces

$$\frac{1}{t}\frac{\partial^2}{\partial t^2}\left(tM(\psi)(x,vt)\right) = \frac{1}{t^2}\frac{\partial}{\partial t}\left(t^2\,\frac{\partial M(\psi)(x,vt)}{\partial t}\right) = v^2\Delta M(\psi)(x,vt),$$

es decir,

$$\frac{\partial^2}{\partial t^2}\left(tM(\psi)(x,vt)\right) = v^2\Delta\big(tM(\psi)(x,vt)\big),$$

como queríamos probar.

La fórmula (13.18) muestra que la constante v admite la misma interpretación que en el caso unidimensional, es decir, se trata de la velocidad de la onda, o la velocidad a la que se desplazan las perturbaciones iniciales. Una diferencia importante es que en la solución tridimensional sólo aparecen integrales sobre los puntos a una distancia vt del punto x, y no sobre los puntos a distancia menor o igual que vt. Esto implica que las condiciones iniciales no pueden causar efectos permanentes. Si suponemos que tienen soporte compacto, para todo punto x existe un instante t a partir del cual $u(x,t)$ se anula.

Vamos a calcular la derivada que aparece en (13.18). Se trata de

$$\begin{aligned}\frac{\partial}{\partial t}\frac{t}{4\pi}\int_{\|\xi\|=1}\phi(x+vt\xi)\,d\sigma(\xi) &= \frac{1}{4\pi}\int_{\|\xi\|=1}\phi(x+vt\xi)\,d\sigma(\xi)\\ &+ \frac{vt}{4\pi}\int_{\|\xi\|=1}\nabla\phi(x+vt\xi)\xi\,d\sigma(\xi)\end{aligned}$$

$$= \frac{1}{4\pi v^2t^2}\int_{\|y-x\|=vt}\phi(y)\,d\sigma(y) + \frac{1}{4\pi v^2 t}\int_{\|y-x\|=vt}\nabla\phi(y)\frac{y-x}{t}\,d\sigma(y).$$

Por consiguiente, (13.18) equivale a

$$u(x,t) = \frac{1}{4\pi v^2t^2}\int_{\|y-x\|=vt}\big(\phi(y)+\nabla\phi(y)(y-x)+t\psi(y)\big)\,d\sigma(y).$$

El caso bidimensional Aunque la ecuación de ondas bidimensional no nos va a hacer falta en las consideraciones posteriores sobre el electromagnetismo, lo cierto es que los cálculos de la sección anterior nos permiten resolverla rápidamente, y la ecuación tiene interés en sí misma. Sus soluciones describen las vibraciones de una membrana, o las vibraciones de la superficie del agua provocadas por la caída de un objeto. Para resolver el problema

$$\left.\begin{array}{rcl} \dfrac{\partial^2 u}{\partial t^2} = v^2\left(\dfrac{\partial^2 u}{\partial x^2}+\dfrac{\partial^2 u}{\partial y^2}\right) & & \\ u(x,y,0) & = & \phi(x,y) \\ \dfrac{\partial u}{\partial t}(x,y,0) & = & \psi(x,y) \end{array}\right\}$$

basta notar que si u es una solución entonces la función $\overline{u}(x,y,z,t)=u(x,y,t)$ es solución del problema tridimensional determinado por las condiciones iniciales $\overline{\phi}(x,y,z)=\phi(x,y)$, $\overline{\psi}(x,y,z)=\psi(x,y)$ —lo que prueba la unicidad— así como que una solución cualquiera $\overline{u}$ del problema tridimensional (con las condiciones iniciales $\overline{\phi}$ y $\overline{\psi}$) no depende de la variable z (pues la función ,$\overline{u}(x,y,z+k,t)$ es solución del mismo problema), luego determina una solución $u(x,y,t)=\overline{u}(x,y,0,t)$ del problema bidimensional.

Para trabajar con la fórmula (13.18) conviene cambiar la notación a $x=(x_1,x_2,x_3)$ identificando los puntos de $\mathbb{R}^2$ con los de la forma $(x_1,x_2,0)$. Entonces

$$u(x_1,x_2,t)=\frac{\partial}{\partial t}\left(\frac{1}{4\pi v^2 t}\int_{\|y-x\|=vt}\overline{\phi}(y)\,d\sigma(y)\right)+\frac{1}{4\pi v^2 t}\int_{\|y-x\|=vt}\overline{\psi}(y)\,d\sigma(y)$$

Puesto que $\overline{\phi}(y_1,y_2,y_3)=\overline{\phi}(y_1,y_2,-y_3)$, la primera integral es el doble de la integral restringida a la semiesfera $y_3>0$. Lo mismo se aplica a la segunda integral. Una carta de dicha semiesfera es

$$X(y_1,y_2)=\sqrt{v^2t^2-(y_1-x_1)^2-(y_2-x_2)^2}=\sqrt{v^2t^2-\|x-y\|^2},$$

donde en la última expresión consideramos x, $y\in\mathbb{R}^2$. Para esta carta

$$d\sigma=\frac{vt}{\sqrt{v^2t^2-\|x-y\|^2}}\,dy_1dy_2,$$

luego

$$\begin{aligned} u(x,t) &= \frac{\partial}{\partial t}\frac{1}{2\pi v}\int_{\|y-x\|<vt}\frac{\phi(y)}{\sqrt{v^2t^2-\|x-y\|^2}}\,dm(y) \\ &+ \frac{1}{2\pi v}\int_{\|y-x\|<vt}\frac{\psi(y)}{\sqrt{v^2t^2-\|x-y\|^2}}\,dm(y). \end{aligned}$$

Esta expresión muestra un comportamiento que difiere tanto del caso unidimensional como del tridimensional. Suponiendo condiciones iniciales con soporte compacto, en el caso tridimensional la función u se anula pasado un tiempo, en el caso bidimensional se va atenuando pero nunca llega a anularse, mientras que en el caso unidimensional la situación depende de las condiciones iniciales.

La ecuación de ondas no homogénea Los resultados anteriores permiten estudiar el comportamiento del campo electromagnético en el vacío (en ausencia de cargas), pues en tal caso las ecuaciones de ondas que hemos obtenido para E, H y sus potenciales son homogéneas. Nos ocupamos ahora del caso general, es decir, del problema tridimensional

$$\left.\begin{aligned} \Box u(x,t) &= w(x,t) \\ u(x,0) &= \phi(x) \\ \frac{\partial u}{\partial t}(x,0) &= \psi(x) \end{aligned}\right\}$$

Ante todo observamos que si existe solución es única, ya que si u_1 y u_2 son soluciones del problema entonces $u_1 - u_2$ es solución del problema homogéneo con condiciones iniciales nulas, luego $u_1 - u_2 = 0$. Para probar la existencia de solución podemos suponer $\phi = \psi = 0$, ya que una solución con otras condiciones iniciales se obtiene añadiendo a la de este caso una solución del problema homogéneo correspondiente.

Para cada $s > 0$ consideremos el problema auxiliar

$$\left.\begin{aligned} \Box u_s(x,t) &= 0 \\ u_s(x,0) &= 0 \\ \frac{\partial u_s}{\partial t}(x,0) &= w(x,s) \end{aligned}\right\}$$

que nos da la onda que aparecería si a partir de una situación de reposo $\phi = 0$ perturbáramos el medio con unas velocidades dadas por $w(x,s)$. Su solución es

$$u_s(x,t) = \frac{1}{4\pi v^2 t}\int_{\|y-x\|=vt} w(y,s)\,d\sigma(y).$$

Para $t > s$ definimos $\overline{u}(x,t,s) = u_s(x,t-s)$, que tiene la misma interpretación salvo que la perturbación aparece en el instante $t = s$ en lugar de en $t = 0$. Entonces

$$\begin{aligned} \frac{\partial^2\overline{u}}{\partial t^2} &= \frac{\partial^2 u_s}{\partial t^2} = v^2\Delta u_s(x,t-s) = v^2\Delta\overline{u}(x,t,s) \\ \overline{u}(x,s,s) &= u_s(x,0) = 0 \\ \frac{\partial\overline{u}}{\partial t}(x,s,s) &= \frac{\partial u_s}{\partial t}(x,0) = w(x,s) \end{aligned}$$

Finalmente definimos

$$u(x,t) = \int_0^t \overline{u}(x,t,s)\,ds = \frac{1}{4\pi v^2}\int_0^t\left(\frac{1}{t-s}\int_{\|y-x\|=v(t-s)} w(y,s)\,d\sigma(y)\right)ds,$$

que representa la acumulación de los efectos de todas las perturbaciones que han aparecido hasta el instante t. Vamos a probar que u es la solución del problema no homogéneo. Para derivar u respecto de t definimos

$$u(x,t_1,t_2) = \int_0^{t_1} \overline{u}(x,t_2,s)\,ds$$

y aplicamos la regla de la cadena. El resultado es

$$\begin{aligned}
\frac{\partial u}{\partial t} &= \overline{u}(x,t,t) + \int_0^t \frac{\partial \overline{u}}{\partial t}(x,t,s)\,ds = \int_0^t \frac{\partial \overline{u}}{\partial t}(x,t,s)\,ds,\\
\frac{\partial^2 u}{\partial t^2} &= \frac{\partial \overline{u}}{\partial t}(x,t,t) + \int_0^t \frac{\partial^2 \overline{u}}{\partial t^2}(x,t,s)\,ds\\
&= w(x,t) + \int_0^t v^2 \Delta \overline{u}(x,t,s)\,ds = w(x,t) + v^2 \Delta u(x,t).
\end{aligned}$$

Evidentemente u cumple las condiciones iniciales. Podemos expresar la solución de forma más simple. Mediante el cambio de variable $s' = v(t-s)$ queda

$$\begin{aligned}
u(x,t) &= \frac{1}{4\pi v^2} \int_0^{vt} \int_{\|y-x\|=s} \frac{w(y, t-s/v)}{s}\, d\sigma(y) ds\\
&= \frac{1}{4\pi v^2} \int_{\|y-x\|<vt} \frac{w(y, t-\|y-x\|/v)}{\|y-x\|}\, dm(y).
\end{aligned}$$

No olvidemos que a esta función hay que sumarle la de una ecuación homogénea en caso de que las condiciones iniciales no sean nulas.

13.5 Soluciones de las ecuaciones de Maxwell

Con los resultados de la sección anterior podemos resolver las ecuaciones (13.10), (13.11), (13.12) y (13.13). Por ejemplo, la primera nos permite expresar el potencial eléctrico V en términos de la densidad de carga ρ. Se cumple

$$V(x,t) = \frac{1}{4\pi\epsilon} \int_{\|y-x\|<ct} \frac{\rho(y, t-\|y-x\|/c)}{\|y-x\|}\, dm(y). \tag{13.19}$$

En principio falta sumar una solución de la ecuación de ondas homogénea correspondiente a las condiciones iniciales de V, pero admitiendo que ρ tiene soporte compacto sabemos que dicha solución forma un frente de onda que se aleja a la velocidad de la luz y tras su paso no deja efecto alguno. En los problemas concernientes a regiones del espacio pequeñas en comparación con la velocidad de la luz podemos prescindir de esta parte.

La fórmula anterior es similar a la del potencial electrostático (potencial newtoniano) salvo por el hecho de que para calcular la influencia en un punto x de la carga situada en un punto y no se tiene en cuenta la carga actual, sino la que había en un instante anterior, el correspondiente al tiempo que tarda la luz en ir de y a x. Por ello las funciones de este tipo se llaman *potenciales retardados.* Vemos, pues, que los efectos en el potencial de una variación de la distribución de las cargas en una región no se hacen notar instantáneamente en todo punto, sino que se transmiten al espacio circundante a la velocidad de la luz. Si las distancias a considerar son pequeñas en comparación con la velocidad de la luz, podemos suponer que la integral se calcula sobre todo $\mathbb{R}^3$ (o sobre todo el soporte de ρ) e incluso prescindir del término $\|y-x\|/c$, con

lo que obtenemos exactamente el potencial electrostático. Por consiguiente, el potencial eléctrico puede suponerse newtoniano en todos los problemas que no involucran distancias astronómicas (lo que supone considerar que las variaciones de la distribución de las cargas alteran instantáneamente el potencial).

Las mismas consideraciones valen para el potencial magnético, que según la ecuación (13.11) viene dado por

$$A(x,t) = \frac{\mu}{4\pi} \int_{\|y-x\|<ct} \frac{\vec{\imath}(y, t - \|y-x\|/c)}{\|y-x\|} \, dm(y). \tag{13.20}$$

Así mismo, las ecuaciones (13.12) y (13.13) proporcionan expresiones similares para E y H.

Ejemplo Vamos a calcular el campo electromagnético creado por una corriente alterna. Primero veremos algunas generalidades sobre este tipo de corriente. En principio una corriente alterna es una corriente que cambia de sentido periódicamente, de modo que pasa de un valor máximo I_0 a un valor mínimo $-I_0$ en un tiempo $T/2$, de modo que al cabo de un tiempo T toma el mismo valor I_0. El caso más simple consiste en suponer que la variación es sinusoidal, es decir, que la intensidad en un tiempo t viene dada por

$$I(t) = I_0 \cos(\omega t + \phi_0), \tag{13.21}$$

donde $\omega = 2\pi/T$. El tiempo T se llama *período* de oscilación y es el tiempo que I tarda en volver al mismo estado. El inverso $1/T$ se llama *frecuencia* de la oscilación y es el número de oscilaciones que se producen por unidad de tiempo. Su unidad (1/segundo) se llama *hercio.* La constante ϕ_0 recibe el nombre de *ángulo de fase.* Eligiendo el origen de tiempos podemos suponer $\phi_0 = 0$. Todos estos conceptos se aplican a cualquier magnitud que varíe sinusoidalmente en el tiempo. En general, si $A\cos(\omega t + \phi_0)$ y $B\cos(\omega t + \phi_1)$ son dos magnitudes sinusoidales con el mismo período, se dice que presentan un *desfase* de $\phi_1 - \phi_0$ radianes. Cuando el desfase es nulo se dice que ambas están *en fase.* Cuando digamos que dos magnitudes sinusoidales están *desfasadas,* sin indicar el desfase, se entenderá que éste es de $\pi/2$ radianes. Dos magnitudes con este desfase pueden representarse por $A\cos(\omega t + \phi_0)$ y $B\,\mathrm{sen}(\omega t + \phi_0)$.

Volviendo a la intensidad (13.21), en principio se trata de la intensidad en un punto dado del cable eléctrico. Ésta no tiene por qué ser la misma en todos los puntos al mismo tiempo, pero ahora no vamos a entrar en ello. Concentrémonos en buscar la mejor forma de tratar con magnitudes sinusoidales.

Resulta que lo más adecuado es sustituir los cosenos por exponenciales complejas. En efecto, la fórmula anterior (con $\phi_0 = 0$) puede escribirse como

$$I(t) = \mathrm{Re}(I_0 e^{i\omega t}).$$

En lo sucesivo escribiremos simplemente $I(t) = I_0 e^{i\omega t}$, es decir, a partir de ahora $I(t)$ no será la intensidad, sino la función compleja cuya parte real es la intensidad y cuya parte imaginaria es la intensidad desfasada (en $-\pi/2$ radianes). Lo mismo valdrá para todas las magnitudes sinusoidales con las que

trabajemos. El teorema 12.9 garantiza que si derivamos la función compleja asociada a una magnitud sinusoidal obtenemos la función compleja asociada a su derivada.

Vamos a estudiar el modelo más simple de corriente alterna. Supongamos que el conductor es un segmento muy pequeño de longitud h cuyo centro está en el origen de coordenadas y se extiende en la dirección del eje z. En tal caso podemos suponer que la intensidad es la misma en cada instante a lo largo de todo el circuito y viene dada por $I = I_0 e^{i\omega t}$. La fórmula (13.20) muestra que el potencial magnético en cada punto tiene la dirección del eje z. Además si el conductor tiene sección k podemos hacer $dm = k\,ds$, donde ds es el elemento de longitud del mismo y así

$$A_z = \frac{\mu}{4\pi}\int_\Omega \frac{\vec{\imath}(y, t - \|y-x\|/c)}{\|y-x\|}\,dm(y) = \frac{\mu}{4\pi}\int_\gamma \frac{I(y, t-\|y-x\|/c)}{\|y-x\|}\,ds(y),$$

donde $\gamma(s) = (0,0,s)$, para $s \in [-h/2, h/2]$. Si nos centramos en puntos alejados del conductor, la distancia $\|y-x\|$ es equiparable a $\|x\|$, con lo que el integrando se vuelve constante y resulta

$$A_z \approx \frac{\mu h}{4\pi}\frac{I(t-\|x\|/c)}{\|x\|} = \frac{\mu h I_0}{4\pi\|x\|}e^{i\omega t}e^{-i\kappa\|x\|},$$

donde $\kappa = \omega/c$.

Es importante observar que al sustituir I por su valor estamos suponiendo que el circuito ya oscilaba en el instante $t - \|x\|/c$, o lo que es lo mismo, que un rayo de luz ha tenido tiempo de llegar desde el origen hasta el punto x en el tiempo que dura la oscilación. Si, por ejemplo, suponemos que la fórmula $I = I_0 e^{i\omega t}$ es válida para $t > 0$ y que antes no había corriente, entonces la fórmula anterior vale para $t > \|x\|/c$ y antes el potencial es nulo.

Observamos que A oscila con periodo T, al igual que la corriente, sólo que entre puntos distintos existe un desfase determinado por el factor $e^{-i\kappa\|x\|}$. No podemos calcular el potencial eléctrico V a partir de la fórmula (13.19), pues ello nos obligaría a estudiar las variaciones locales de la densidad de carga. En su lugar razonamos como sigue: el potencial eléctrico debe oscilar también con periodo T, es decir, ha de ser de la forma $V = f(x)e^{i(\omega t + \phi_0(x))}$, luego

$$\frac{\partial V}{\partial t} = i\omega f(x)e^{i(\omega t+\phi_0(x))} = i\omega V.$$

Por el mismo razonamiento

$$\frac{\partial A}{\partial t} = i\omega A. \tag{13.22}$$

La condición de Lorentz (13.4), que hemos impuesto a los potenciales, nos da ahora que

$$V = \frac{i}{\omega\mu\epsilon}\,\mathrm{div}\,A = \frac{ic^2}{\omega}\,\mathrm{div}\,A = \frac{i\omega}{\kappa^2}\,\mathrm{div}\,A. \tag{13.23}$$

A partir de aquí los cálculos se simplifican mucho si pasamos a coordenadas esféricas. Notemos que el tercer vector de la base canónica es

$$\nabla z = \nabla(r\cos\theta) = \cos\theta\, v_r - \operatorname{sen}\theta\, v_\theta.$$

Por consiguiente

$$\begin{aligned} A_r &= A_z\cos\theta = \frac{\mu h I_0}{4\pi r}e^{i\omega t}e^{-i\kappa r}\cos\theta, \\ A_\theta &= -A_z\operatorname{sen}\theta = \frac{\mu h I_0}{4\pi r}e^{i\omega t}e^{-i\kappa r}\operatorname{sen}\theta, \\ A_\phi &= 0. \end{aligned}$$

Ahora es fácil calcular

$$H = \frac{1}{\mu}\operatorname{rot} A = \frac{1}{\mu r^2\operatorname{sen}\theta}\begin{vmatrix} v_r & r v_\theta & r\operatorname{sen}\theta\, v_\phi \\ \frac{\partial}{\partial r} & \frac{\partial}{\partial \theta} & 0 \\ A_r & rA_\theta & 0 \end{vmatrix},$$

lo que nos lleva a $H_r = H_\theta = 0$ y

$$H_\phi = \frac{I_0 h}{4\pi}e^{i\omega t}e^{-i\kappa r}\left(\frac{i\kappa}{r} + \frac{1}{r^2}\right)\operatorname{sen}\theta.$$

Por otra parte, (13.23) nos lleva a

$$V = \frac{\mu I_0 h}{4\pi}e^{i\omega t}e^{-i\kappa r}\left(\frac{\omega}{\kappa r} - \frac{i\omega}{\kappa^2 r^2}\right)\cos\theta.$$

Esto juntamente con (13.22) nos permite calcular

$$E = -\nabla V - \frac{\partial A}{\partial t} = -\nabla V - i\omega A.$$

tras un cálculo rutinario obtenemos

$$\begin{aligned} E_r &= \frac{I_0 h}{4\pi}e^{i\omega t}e^{-i\kappa r}\left(\frac{2\eta}{r^2} + \frac{2}{i\omega\epsilon r^3}\right)\cos\theta, \\ E_\theta &= \frac{I_0 h}{4\pi}e^{i\omega t}e^{-i\kappa r}\left(\frac{i\omega\mu}{r} + \frac{\eta}{r^2} + \frac{1}{i\omega\epsilon r^3}\right)\operatorname{sen}\theta, \\ E_\phi &= 0, \end{aligned}$$

donde $\eta = \sqrt{\mu/\epsilon}$ es una constante cuyas dimensiones son las de una resistencia y en el vacío vale $\eta_0 \approx 120\pi$ ohmios.

Para puntos alejados del circuito emisor los términos en $1/r^2$ y $1/r^3$ son despreciables frente a los términos en $1/r$, con lo cual una buena aproximación al campo electromagnético creado por el mismo es

$$H_\phi \approx \frac{i\kappa I_0 h}{4\pi r}e^{i\omega t}e^{-i\kappa r}\operatorname{sen}\theta, \tag{13.24}$$

$$E_\theta \approx \frac{i\omega\mu I_0 h}{4\pi r}e^{i\omega t}e^{-i\kappa r}\operatorname{sen}\theta, \tag{13.25}$$

con las componentes restantes nulas o muy pequeñas. Geométricamente, esto significa que los campos E y H son perpendiculares entre sí y perpendiculares al radio que los une con el circuito emisor. Notemos además que $E_\theta = \eta H_\phi$, luego los campos están en fase.

En resumen, el circuito está rodeado de un campo electromagnético que ocupa un volumen esférico cuyo radio se extiende a la velocidad de la luz. En cada punto de este volumen el campo oscila con período T, la intensidad máxima de los campos varía entre 0 en el eje Z y un valor máximo en el plano XY. Fijada una recta que pase por el origen, la intensidad máxima de los campos disminuye en proporción inversa a la distancia. Además la oscilación está desfasada en κr radianes, de modo que los campos en dos puntos de la recta están en fase si y sólo si su distancia es un múltiplo de $\lambda = 2\pi/\kappa = cT$ metros. Este valor se llama *longitud de onda.*

Observemos que (con las aproximaciones que hemos hecho) el vector de Poynting es siempre perpendicular a las esferas de centro en el origen y apunta siempre hacia fuera de las mismas. Por lo tanto, aunque es oscilante, su oscilación no es sinusoidal, sino que depende de un seno al cuadrado. La intensidad (el módulo) del vector de Poynting en un punto y en un instante dado es

$$P = \|E\|\,\|H\| = \frac{\kappa^2 I_0^2 h^2}{16\pi^2 r^2}\,\eta \operatorname{sen}^2\theta \operatorname{sen}^2(\omega t - \kappa r) = \frac{I_0^2 h^2 \eta}{4\lambda^2 r^2} \operatorname{sen}^2\theta \operatorname{sen}^2(\omega t - \kappa r),$$

de modo que P oscila entre 0 y un valor máximo con período $T/2$. Para una magnitud que oscila de este modo tiene sentido calcular su valor medio, definido como

$$P_m(r,\theta) = \frac{2}{T}\int_0^{T/2} P(r,\theta,t)\,dt.$$

Concretamente

$$P_m(r,\theta) = \frac{I_0^2 h^2 \eta}{4\lambda^2 r^2} \operatorname{sen}^2\theta\,\frac{1}{\pi}\int_{-\kappa r}^{\pi-\kappa r} \operatorname{sen}^2 x\,dx = \frac{I_0^2 h^2 \eta}{8\lambda^2 r^2} \operatorname{sen}^2\theta \quad \text{vatios/metro}^2.$$

La potencia irradiada, o cantidad de energía electromagnética que atraviesa por segundo una esfera de centro en el conductor es el flujo del vector de Poynting a través de la esfera, es decir,

$$\begin{aligned} W(r,t) &= \int_{S_r} P\,d\sigma = \int_{S_r} P(r,\theta,t)\,r^2 \operatorname{sen}\theta\,d\theta\wedge d\phi \\ &= \frac{I_0^2 h^2 \eta}{4\lambda^2} \operatorname{sen}^2(\omega t - \kappa r)\int_0^{2\pi}\int_0^{\pi} \operatorname{sen}^3\theta\,d\theta d\phi \\ &= \frac{2\pi I_0^2 h^2 \eta}{3\lambda^2} \operatorname{sen}^2(\omega t - \kappa r) \quad \text{vatios.} \end{aligned}$$

La potencia W depende de r únicamente en el desfase entre puntos situados a distancias distintas del circuito. Sin embargo la potencia media no depende de r, pues se comprueba fácilmente que vale

$$W_m = \frac{\pi I_0^2 h^2 \eta}{3\lambda^2} \approx 40\pi^2 I_0^2 \frac{h^2}{\lambda^2} \quad \text{vatios.}$$

Ésta es la energía que irradia por segundo el circuito. (En un tiempo nT, con n natural, la energía irradiada será nW_m. Si n no es natural el valor nW_m es aproximado.)

Hemos visto que el campo determinado por el vector de Poynting es radial, pero en regiones alejadas de la fuente de radiación podemos considerarlo paralelo. Escogiendo adecuadamente el sistema de referencia podemos suponer que tiene la dirección del eje X, así como que los campos H y E tienen, respectivamente, la dirección de los ejes Y y Z. Entonces las ecuaciones (13.24) y (13.25) son aproximadamente

$$H = \frac{I_0 h \operatorname{sen}\theta}{2\lambda x} \operatorname{sen}(\omega t - \kappa x)e_2, \qquad E = \eta\frac{I_0 h \operatorname{sen}\theta}{2\lambda x} \operatorname{sen}(\omega t - \kappa x)e_3,$$

donde ahora θ es constante. Si la longitud de onda λ es pequeña, una variación moderada de x altera en muy poco el factor $2\lambda x$, con lo que la intensidad de la onda se puede considerar constante y queda

$$H = A \operatorname{sen}(\omega t - \kappa x)e_2, \qquad E = \eta A \operatorname{sen}(\omega t - \kappa x)e_3,$$

para una cierta constante A. Ésta es la forma más simple que puede adoptar el campo electromagnético en el vacío. Es lo que se llama una onda *plana* (porque el frente de ondas es plano), *transversal* (porque los campos varían perpendicularmente a la dirección de avance), *monocromática* (porque la variación en cada punto es sinusoidal) y *polarizada* (porque E y H varían siempre en la misma dirección).

Las ondas electromagnéticas monocromáticas se clasifican por su longitud de onda:

Ondas de radio	30 Km $> \lambda >$ 400 μm
Rayos infrarrojos	400 μm $> \lambda >$ 0,8 μm
Luz (visible)	0,8 μm $> \lambda >$ 0,4 μm
Rayos ultravioletas	0,4 μm $> \lambda >$ 120 Å
Rayos X	120 Å $> \lambda >$ 0,05 Å
Rayos γ	0,05 Å $> \lambda$

Hemos usado el *micrómetro* o *micra,* abreviado μm, igual a una milésima de milímetro, y el *Angstrom,* abreviado Å, igual a 10^{-10}m. Las ondas monocromáticas de longitud entre 0,8 y 0,4 micras son casos particulares de lo que comúnmente llamamos "luz". El ojo humano percibe la longitud de onda en forma de color. Las longitudes cercanas a 0,8 micras corresponden a la luz roja, mientras que las cercanas a 0,4 micras corresponden a la luz violeta, pasando por toda la gama del arco iris. Los colores que no aparecen en el arco iris, como el marrón, corresponden a superposiciones de luz de distintas longitudes de onda. ■

Bibliografía

[1] Borden, R.S. *A course in advanced calculus.* North Holand, Amsterdam, 1983.

[2] Corwin, L.J. Szczarba, R.H. *Multivariable calculus.* Marcel Dekker, New York, 1982.

[3] Do Carmo, M.P. *Geometría diferencial de curvas y superficies.* Alianza, Madrid, 1990.

[4] Elsgoltz, L. *Ecuaciones diferenciales y cálculo variacional.* Ed. Mir, Moscú, 1977.

[5] Greub, W., Halperin, S., Vanstone, R. *Connections, curvature, and cohomology* Vol. 1. Academic Press, New York, 1972.

[6] Hu, S.T. *Differentiable manifolds.* Holt, Rinehart and Winston, New York, 1969.

[7] John, F. *Partial differential equations.* Springer, New York, 1982.

[8] Lang, S. *Differential Manifolds.* Addison Wesley, Reading, Mass. 1972

[9] Ramo, S. Whinnery, J.R., van Duzer, T. *Campos y ondas.* Pirámide, Madrid, 1965.

[10] Rudin, W. *Análisis real y complejo.* Mc. Graw Hill, Madrid, 1988.

[11] Santaló, L. A. *Vectores y tensores.* Ed. Universitaria de Buenos Aires, 1968.

Índice de Materias